

IDENTIFICATION OF POTENTIAL ANTIBIOFILM
HIT COMPOUNDS FROM TWO AFRICAN
NATURAL PRODUCT DATABASE AGAINST
MULTI-DRUG RESISTANT STAPHYLOCOCCUS
AUREUS: AN *IN SILICO* STUDY

ILORI TOSIN LYDIA (4163974)

PROF SAMUEL EGIEYEH (Supervisor)

A thesis submitted to the School of Pharmacy, Faculty of Natural
Sciences, University of the Western Cape

in partial fulfilment of the requirements for the degree of

Masters in Pharmacy

2023

ABSTRACT

One of the crucial ways by which *Staphylococcus aureus* develops resistance to antibiotics is biofilm formation, a protective mechanism involving extracellular polymeric substance (EPS) matrix that shields microorganisms from the effects of antibiotics, mechanical forces, pH, and host immune responses. While some encouraging results point to the possible use of FDA-approved medications against biofilms, more research is needed due to sporadic and patchy data. The complex chemical diversity of natural compounds makes them a reservoir of bioactive molecules for drug discovery. This study seeks to identify effective potential antibiofilm compounds from a query dataset compiled from two African natural product databases (SANCDb and AfroDb).

A database of known antibiofilm compounds was created from ChEMBL, PubChem, and other related databases while a query dataset of natural products was compiled for this study. The ligand similarity (LS) searches were unable to unequivocally identify distinct differences in the molecular structures and functional group moiety of the active and inactive compounds. The flexophore similarity metric detected correlations between the query dataset and the known antibiofilm molecules. Using a machine learning approach, the Random Forest (RF) predictive model displayed better superior accuracy in predicting the antibiofilm bioactivity of natural compounds in the query database. Consensus scoring of compounds identified from LS searches and the RF predictive model was done to select hit compounds for docking. The analysis of docking scores revealed that the CNP0160461 and CNP0037371 exhibit the strongest binding with identified *Staphylococcus aureus* biofilm-associated proteins.

TABLE OF CONTENTS

ABSTRACT	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF APPENDICES	xi
LIST OF ABBREVIATIONS	xii
ACKNOWLEDGEMENT	xiv
Chapter One.....	1
Introduction.....	1
1.1 Background Studies.....	1
1.2 Research problem	2
1.3 Significance of the study	3
1.4 Aim	3
1.6 Thesis outline.....	4
Chapter Two.....	7
Literature review	7
2.1 <i>Staphylococcus aureus</i> infections.....	7
2.2.1 Overview of microbial biofilm formation	9
2.2.2 Developmental stages of biofilm formation	11
2.2.3 Clinical burden constituted by Staphylococcal biofilms	13
2.3 Epidemiology.....	14
2.3 Current approaches, success, and challenges to discovering antibiofilm compounds.....	15
2.3.1 Nanotechnology	16
2.3.2 Quorum sensing	16
2.3.3 Use of Biofilm dispersal strategy.....	17
2.3.4 Bacteriophages.....	18
2.3.5 Biosurfactants	19
2.4 Natural compounds as a potential source of anti-biofilm compounds	21
2.5 The use of computational tools in drug design.....	23
2.5.1 Structure-based computer-aided drug discovery	25
2.5.2 Ligand-based computer-aided drug discovery.....	26

2.5.3 Molecular similarity search.....	27
2.5.4 Machine learning	28
2.6 Conclusion.....	29
Chapter Three.....	31
Ligand similarity approach to discovery of potential antibiofilm hit compounds from two African natural product databases against multidrug resistant <i>Staphylococcus aureus</i>	31
3.1 Introduction	31
3.2 Method.....	33
3.2.1 Collection of active and inactive antibiofilm compounds against <i>Staphylococcus aureus</i>	34
3.2.2 Query databases	35
3.2.3 Retrieval of SMILES structures of datasets and generation of compound structures.....	35
3.2.4 Data characterization	35
3.2.5 Calculation of compound properties and descriptors.....	36
3.2.6 Similarity charts and Scaffold analysis	37
3.2.7 Flexophore similarity.....	38
3.3 Results and Discussion	38
3.3.1 Data collection and characterization.....	38
3.3.2 Calculation of properties and descriptors	39
3.3.3 Similarity charts.....	44
3.3.4 Analysing scaffold.....	51
3.3.5 Flexophore similarity search by comparing Active antibiofilm compounds and Query dataset	55
3.4 Conclusion.....	58
Chapter Four.....	59
Building a predictive model using a machine learning approach	59
4.1 Introduction	59
4.2 Materials and Methods	60
4.2.1 Data	61
4.2.2 Machine learning algorithms	61
4.2.3 Dataset pre-processing and calculation of molecular descriptors and molecular fingerprints.....	62
4.2.4 Class Imbalance and cost-sensitive classification.....	62

4.2.5 Selection of descriptors and features	62
4.2.6 Training and Evaluation of performance of antibiofilm predictive models	63
4.3 Results and Discussion	67
4.4 Conclusion	70
Chapter five	72
Consensus scoring for compounds from flexophore similarity search and Random Forest predicted model	72
5.1 Introduction	72
5.2 Methods and Materials	72
5.3 Results and Discussion	74
5.4 Conclusion	77
Chapter Six	79
Reverse molecular docking of top ranking consensus-scored compounds with antibiofilm activity	79
6.1 Introduction	79
6.2 Methods and Materials	79
6.2.1 Ligand preparation	80
6.2.2 Protein selection	80
6.2.3 Identifying binding sites in protein targets	82
6.2.4 Grid generation and Docking	83
6.3 Results and Discussion	84
6.4 Conclusion	86
Chapter Seven	87
Conclusion, limitations and recommendations	87
7.1 Summary of findings	87
7.2 Limitations	89
7.3 Recommendations and Conclusion	90
Appendices	91
References	97

LIST OF TABLES

Table 1 Value of the Accuracy and ROC curve of the models	68
Table 2 Summary of interaction analysis for <i>Staphylococcus aureus</i> biofilm-associated proteins and identified hits from consensus scoring	86



LIST OF FIGURES

- Figure 2.1** Developmental stages of *S. aureus* biofilm (Adapted from Alves Carneiro et al., 2020) 13
- Figure 3.1(a)** Box plot of cLogP of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively. 40
- Figure 3.1(b)** Box plot of H-Donors of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively. 40
- Figure 3.1(c)** Box plot of Aromatic atoms of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively. 41
- Figure 3.1(d)** Box plot of Hetero-rings of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively 42
- Figure 3.1(e)** Box plot of Saturated Rings of active and inactive antibiofilm compounds against *Staphylococcus aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively 43
- Figure 3.2(a)** Resulting similarity tree for both active and inactive antibiofilm compounds using molecular fragment fingerprint. Active antibiofilm compounds denoted with red dots and non-active are blue dots. Similar neighbour compounds are connected with a connecting line to form a cluster. 46
- Figure 3.2(b)** Resulting similarity tree for both active and inactive antibiofilm compounds using flexophore molecular fingerprint. Active and inactive antibiofilm compounds are denoted in red dots and blue dots respectively. Similar neighbour compounds with similar flexophore are connected with a connecting line to form a cluster. 46
- Figure 3.3(a)** Analysis of neighbour tree A showing the core structure and substituents with their corresponding activity class 47
- Figure 3.3(b)** Analysis of neighbour tree A cont'd. showing the core structure and substituents with their corresponding activity class 47
- Figure 3.3(c)** Analysis of neighbour tree A cont'd showing the core structure and substituents with their corresponding activity class 48
- Figure 3.3(d)** Analysis of neighbour tree A cont'd. showing the core structure and substituents with their corresponding activity class 48
- Figure 3.3(e)** Analysis of neighbour tree B showing the core structure and substituents with their corresponding activity class 49
- Figure 3.3(f)** Analysis of neighbour tree B cont'd. showing the core structure and substituents with their corresponding activity class 49

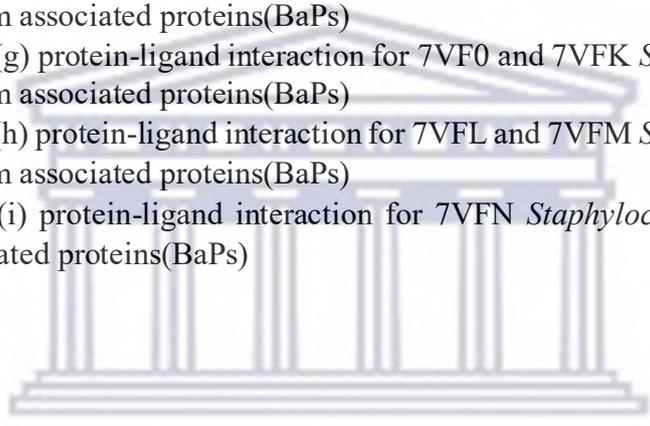
Figure 3.3(g) Analysis of neighbour tree C showing the core structure and substituents with their corresponding activity class	50
Figure 3.3(h) Analysis of neighbour tree C cont'd. showing the core structure and substituents with their corresponding activity class	51
Figure 3.4(a) Murcko scaffolds of Active antibiofilm using Datawarrior. Unique scaffolds that are similar to those in the inactive group are labelled A, B, C, and D. The colours indicate the frequency of the scaffolds, where the blue and red colour represent smallest and largest values respectively	53
Figure 3.4(b) Murcko scaffolds of Inactive antibiofilm using Datawarrior. Unique scaffolds that are similar to those in the active group are labelled A, B, C, and D. The colours indicate the frequency of the scaffolds, where the blue and red colour represent smallest and largest values respectively.	54
Figure 3.4(c) Comparison between scaffolds A, B, C, and D of active and inactive antibiofilm compounds	55
Figure 3.5 Histogram showing the frequency distribution of flexophore similarity score. Total count 43,957 of the query dataset, mean similarity score 0.88926, min 0.850000, Std 0.033, 25% 0.863030, 50% 0.880180, 75% 0.907270, max 1.000000	58
Figure 4.1(b) Screenshot of the KNIME workflow used to build the MLP classifier.	66
Figure 4.1(c) Screenshot of the KNIME workflow used to build the XGBOOST classifier machine-learning model	67
Figure 4.1(d) Screenshot of the KNIME workflow used to build the SVM classifier machine-learning model	67
Figure 4.2 A and C represent ROC-AUC curve of Random forest and XGBOOST with B and D showing overfitting of the respective models with MACC+ FCFP6 fingerprint. Dark blue and Red colour represent active and inactive compounds respectively. ROC-AUC plot evaluates model performance after training	70
Figure 5.1(a) Showing the Distribution of RF-ML prediction confidence. The prediction scaled from '0' to '1', the higher the prediction confidence value, the better the chance antibiofilm property	76
Figure 5.1(b) Histogram showing the distribution of flexophore similarity score. The score ranges from '0.85' to '1', the higher the score the greater the similarity between the natural compound flexophore and the known active antibiofilm compound flexophore	76
Figure 5.1(c) Histogram showing the distribution of average score = (flexophore similarity score + RF ML prediction confidence)/2	77
Figure 5.1(d) Histogram showing the distribution of Z-score normalized average score	77
Figure 6 Showing the glide docking scores of the <i>S. aureus</i> biofilm-associated protein with identified natural compounds from the consensus scoring of results from flexophore similarity chart and random forest predictive model. The coloured bars show how a score compares to others. Longer bars represent	

higher docking scores, shorter bars represent smaller docking scores, and missing values represent no existing docking interaction. 85



LIST OF APPENDICES

Appendix (a) protein-ligand interaction for 2VR3 and 3ASW <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	91
Appendix (b) protein-ligand interaction for 3AT0 and 3AU0 <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	91
Appendix (c) protein-ligand interaction for 3GEU and 4F1Z <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	92
Appendix (d) protein-ligand interaction for 4F20 and 4F24 <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	93
Appendix (e) protein-ligand interaction for 4F27 and 5JQ6 <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	93
Appendix (f) protein-ligand interaction for 7C7R and 7EC1 <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	94
Appendix (g) protein-ligand interaction for 7VF0 and 7VFK <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	95
Appendix (h) protein-ligand interaction for 7VFL and 7VFM <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	95
Appendix (i) protein-ligand interaction for 7VFN <i>Staphylococcus aureus</i> biofilm associated proteins(BaPs)	96



UNIVERSITY of the
WESTERN CAPE

LIST OF ABBREVIATIONS

- ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity
- AfroDB: Database of natural products from African sources
- AUC: Area Under the Curve
- BaPs: Biofilm-associated Proteins
- BS: Biosurfactant
- ChEMBL: Chemical database of bioactive molecules with drug-like properties
- CHPC: Centre for High-Performance Computing
- COVID-19: Coronavirus Disease 2019
- DNA: Deoxyribonucleic Acid
- EPS: Extracellular polymeric substances
- HIV: Human immunodeficiency virus
- HTS: High-throughput screening
- MDRSA: Multidrug Resistant *Staphylococcus aureus*
- ML: Machine Learning
- MLP: Multilayer perceptron
- MRSA: Methicillin-Resistance *Staphylococcus aureus*
- MSCRAMMs: Microbial Surface Components Recognizing Adhesive Matrix Molecules
- NMR: Nuclear Magnetic Resonance
- PDB: Protein Data Bank
- QS: Targeting quorum sensing

QSAR: Quantitative structure–activity relationship

RD: Reverse Docking

RF: Random Forest

ROC: Receiver Operating Characteristic

ROS: Reactive Oxygen Species

SAB: *Staphylococcus aureus* bacteremia

SANCDb: South African Natural Compounds Database

SAR: Structural-Activity Relationship

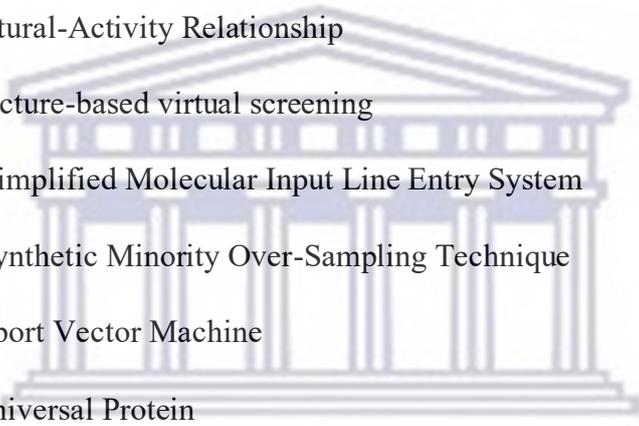
SBVS: Structure-based virtual screening

SMILES: Simplified Molecular Input Line Entry System

SMOTE: Synthetic Minority Over-Sampling Technique

SVM: Support Vector Machine

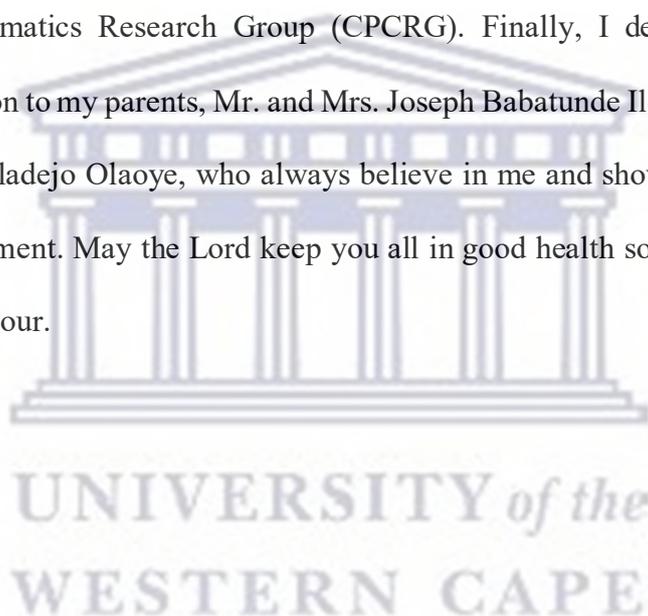
UniProt: Universal Protein



UNIVERSITY *of the*
WESTERN CAPE

ACKNOWLEDGEMENT

I am indeed grateful to God for giving me the grace, strength, and courage to finish this study. To my supervisor, Prof. Samuel Ayodele Egieyeh, thank you for your guidance, contribution, and unwavering support throughout the course of this research. Many thanks to my husband, Dr. Emmanuel Idemudia Ilori, for his amazing love, financial support, encouragement, and for giving me this chance. I would also like to thank the members of the Computational Pharmacology and Cheminformatics Research Group (CPCRG). Finally, I dedicate my unfettered appreciation to my parents, Mr. and Mrs. Joseph Babatunde Ilori and Chief and Mrs. Michael Oladejo Olaoye, who always believe in me and shower me with words of encouragement. May the Lord keep you all in good health so you can eat the fruits of your labour.



Chapter One

Introduction

1.1 Background Studies

Multidrug-resistant *Staphylococcus aureus* (MRSA) infections have become a great threat to public health because they are capable of causing serious and life-threatening infections in humans. Multidrug-resistant (MDR) bacteria are resistant to three or more classes of antimicrobial drugs and are difficult to treat (Jernigan et al., 2020). In terms of drug resistance, *Staphylococcus aureus* has been able to steadily acquire new modes of resistance to nearly all antibiotics (Kakoullis et al., 2021). Resistance is rapidly evolving as a defence mechanism against antibiotics. The use of antibiotics has played an important role in medicine, saving lives, but there is an emergence of almost unavoidable resistance to the available antibiotics. However, while synergistic combination of antibiotics for the treatment of bacterial infections has been successful, there has been some glaring drawbacks as well (Assis et al., 2017). The failure of antibiotic treatments for *Staphylococcus aureus* infections has been attributed to the mechanism of biofilm formation. Biofilm formation offers increased protection from antibiotics, mechanical forces, nutrient scarcity, pH, and host immune response (Divakar et al., 2019). Biofilm have the ability to undergo series of nutrient recycling hence promoting the growth and survival of microorganism within the matrix (Bamford et al., 2023).

1.2 Research problem

Staphylococcus aureus is of major concern because of its ability to cause diverse potentially fatal infections. Contaminated medical devices are viable sources of contact *S. aureus*-associated skin and soft tissue infections. Biofilm formation is the key virulence factor and a key survival strategy for *Staphylococcus aureus*. Biofilms are capable of providing an inactive but constantly changing environment in which the bacterial cells can achieve homeostasis (Tong et al., 2015). Antibiotic resistance is reportedly up to 1000-fold greater in biofilm-bacterial cells, which are able to tolerate significantly higher levels of antibiotics than planktonic bacteria (Penesyanyan et al., 2019). The formation of biofilm by bacteria imposes great challenges on the use of conventional antimicrobials. While some encouraging results point to the possible use of FDA-approved medications against biofilms, more research is necessary as data is still sporadic and patchy (Hawas et al., 2022). Therefore, new, and effective antibiofilm molecules to combat the multidrug-resistance in *Staphylococcus aureus* infections are urgently needed. Due to the fact that natural compounds and their analogues exhibit a vast array of scaffolds and structural complexity, interest in natural products as drug leads is currently resurgent, particularly in the fight against antimicrobial resistance. Because natural compounds and their analogues are characterised by enormous scaffold diversity and structural complexity, interest in natural products as drug leads is currently being revived, particularly for combating antimicrobial resistance (Atanasov et al., 2021). Hence, this study adopted computational approaches such as ligand similarity searches, machine

learning, and molecular docking approaches to identify natural compounds with antibiofilm activity against multidrug-resistant *Staphylococcus aureus*.

1.3 Significance of the study

Infections associated with biofilms are difficult to treat. Prior to now, the conventional laboratory compound testing known as “high throughput screening” was used in the discovery of novel bioactive compounds, but it was time-consuming and inefficient. Also, several compounds had to be synthesised and tested experimentally. Therefore, this study adopted a computational approach that is cost-effective, time-efficient, and reproducible to identify compounds with antibiofilm activity against multidrug-resistant *Staphylococcus aureus* from a natural compound database. The identification of these natural antibiofilm agents and their further drug development will likely reduce the emerging resistance to current antibiotics by *Staphylococcus aureus*. The novel antibiofilm agents can then be used independently or in conjunction with current antimicrobial drugs.

1.4 Aim

This study seeks to identify potential antibiofilm hit compounds from two African natural product databases (AfroDb and SANCDb) against multidrug-resistant *Staphylococcus aureus* using *in-silico* approaches.

1.5 Objectives

The objectives of the study are:

1. To collate and build a database of active and inactive antibiofilm compounds from bioassays in a literature search, collate the biofilm-associated proteins of *Staphylococcus aureus* involved in cellular

aggregation within the biofilm and retrieve compounds from AFRODb and SANCDb. Subsequently, study the properties of compounds with reported antibiofilm properties and conduct a flexophore similarity search between the known active antibiofilm compounds and the query databases.

2. To build an antibiofilm predictive model with a combination of important molecular descriptors and fingerprints of the known active and inactive antibiofilm compounds using a machine learning approach, then use this to predict the antibiofilm activity of the natural compounds from the query databases.
3. To develop a consensus scoring function for the hit compounds identified in objectives 1 and 2 above.
4. To perform molecular docking studies on consensus-scored hit antibiofilm compounds with *Staphylococcus aureus* biofilm-associated proteins to predict the possible mechanism of antibiofilm activity.

1.6 Thesis outline

This thesis consists of seven Chapters in total. **Chapter One** presents the background studies and a general overview of this thesis. This is followed by the identification of the research problem and the study's significance. There's a description of the aim and objectives of the study. The final section of Chapter One is the outline of the subsequent Chapters. **Chapter Two** provides a literature review on *Staphylococcus aureus* infections and a description of developmental stages in biofilm formation. The Chapter discusses the clinical burden constituted by Staphylococcal biofilm-associated infections. In addition, the current

approaches, successes, and challenges to discovering antibiofilm compounds were discussed. It discusses the epidemiology and prevalence of resistance in *Staphylococcus aureus*. Chapter Two further identifies the role of natural compounds as potential antibiofilm in drug discovery. It also discusses the use of computational tools to find new therapies.

Chapter Three provides details of data collection and curation. It explains how the “Simplified Molecular Input Line Entry System (SMILES)” structures of active and inactive antibiofilm compounds and “query databases” retrieved were used to generate their corresponding chemical structures. Ligand similarity searches were conducted by analysing the molecular fragment and flexophore descriptors of known active antibiofilm compounds to identify potential antibiofilm compounds from the query databases. **Chapter Four** involves using machine learning approaches to build antibiofilm predictive models using important molecular features and descriptors of the known actives and inactive compounds. The model with better predictive accuracy was used to predict the antibiofilm activity of natural compounds in the query databases to identify the hits.

In Chapter Five, consensus scoring was used to rank the prediction of potential antibiofilm compounds from ligand similarity searches and Random Forest predictive model. **Chapter Six** utilised reverse docking (RD) to understand the protein-ligand interaction and predict the possible mechanism of action for the antibiofilm activities of the high-ranked consensus-scored compounds when they bind to *Staphylococcus aureus* biofilm-associated proteins. **Chapter Seven**

summarises the major findings of this study based on the objectives in Chapter One. Additionally, it offers the study's limitations and suggestions for future research.



Chapter Two

Literature review

2.1 *Staphylococcus aureus* infections

Staphylococcus aureus is a gram positive bacteria belonging to the genus *Staphylococcus*. It has a diameter of about 0.8µm in diameter, can grow aerobically or anaerobically, and thrives best at 37 °C and pH 7.4 (Guo et al., 2020). On a blood agar plate, they form dense, shiny and round colonies (Gonzalez-Sato et al., 2019). *Staphylococcus aureus* lacks spores or flagella and has a capsule capable of producing yellow pigment and decomposing mannitol. Additionally, it has also been discovered that *S. aureus* tests positive for plasma coagulase, lactose fermentation, and deoxyribonuclease tests (Tayeb-fligelman et al., 2017). *Staphylococcus aureus* has been indicated as one of the most frequent worldwide causes of health problems and death (Cheung et al., 2021). *Staphylococcus aureus* is a significant contributor to both hospital-acquired and community-acquired infections, and it places a heavy burden on the healthcare system. For instance, majority of cases of bone infection (Osteomyelitis) are caused by *S. aureus*. *Staphylococcus aureus* is also capable of infecting orthopaedic implants such as prosthetic joints, external fixtures, fragment implants, etc. (Dasilva et al., 2013). The development of biofilm is crucial in chronic infections (Lister & Horswill, 2014). *Staphylococcus* spp. that develop biofilms are important reservoirs for the spread of ocular infections. Keratitis, conjunctivitis, and endophthalmitis are among the ocular diseases associated with *S. aureus* biofilm (Archer et al., 2011). Chronic wound infections such as diabetic

foot ulcers and venous stasis ulcers have also been linked to *S. aureus* biofilms. Other moderately severe *S. aureus* skin infections, such as furuncles, abscesses, and wound infections are typically not life-threatening but can cause significant morbidity and discomfort. It is known that Staphylococci are the most common cause of infection linked to biofilms. This unique status of Staphylococci among biofilm-associated pathogens is due to the fact that Staphylococci are commonly found commensal bacteria on the human skin and mucous surfaces and those of many other mammals (Otto, 2019). Asymptomatic commensal colonization of *S. aureus* can be considered as an important prerequisite for further infection. The frequent touching and nose picking as well as the distribution that results are thought to be the source of this association. Skin infections may arise from minor scratches on the skin and become invasive if bacterial penetrate through the epithelial protective barrier. Additionally, *Staphylococcus aureus* can be acquired from animals, particularly in the livestock sector where the emergence of livestock-associated Methicillin-resistant *Staphylococcus aureus* (LA-MRSA) has raised serious concerns. A common source of infection in the hospital environment is the contamination of indwelling medical devices. The primary mechanism responsible for this infection route is the ability of *S. aureus* to attach to the devices and to the matrix molecules that cover the devices shortly after insertion, forming a biofilm on the device. Food poisoning is a unique instance of acute *S. aureus* infection that occurs when contaminated foods containing Staphylococcal enterotoxins (SEs) are consumed. *Staphylococcus aureus* can potentially exploit favourable circumstances or initial damage caused by other pathogens in an opportunistic manner. For example, *S. aureus* secondary

infections are frequently the primary cause for death in lung infections that are initiated by a viral infection like the flu (Cheung et al., 2021). In studies involving patients with chronic venous leg ulcers, *S. aureus* was identified as the most frequently isolated bacterium from such wound infections, and *S. aureus*-positive cultures were detected in 88–93.5% of wound infections (Gjodsbol et al., 2006).

2.2.1 Overview of microbial biofilm formation

A biofilm can be referred to as a sessile community of microorganisms characterised by cells that are embedded in a matrix of extracellular polymeric substances (EPS) composed of polysaccharides, proteins, and nucleic acids, resulting in a changed gene expression, protein production, metabolic activity, and growth (Kraranjec et al., 2021). Microorganisms embedded in the biofilm matrix have a low metabolic rate, which explains the antibiotic-resistance properties of biofilms. The biofilm acts as a diffusion barrier to slow down antimicrobial agent infiltration and minimize the concentration of the antibiotic intracellularly as a result of poor biofilm penetration (Archer et al., 2011). The matrix captures and chemically renders antibiotics that somehow manage to find their way into the matrix inactive. Also, efflux pumps and secretion systems actively remove any residual antibiotic from within the biofilm. The biofilm matrix is made up of proteins (e.g., fibrin), essential nutrients, and minerals. The extracellular biofilm matrix contains 1-2% polysaccharides (e.g., alginate), < 1% DNA, < 1% RNA, ions, and 97% water. The EPS matrix (0.2 -1.0µm thick) strengthens the interaction among microorganisms and shields them from mechanical stress or the effects of antibiotics (Sahoo et al., 2021). The composition of the biofilm matrix varies between strains but generally can

contain host factors, polysaccharides, proteins, and extracellular DNA (Montanaro et al., 2011). In a nutrient-deficient condition such as starvation, a good survival strategy is biofilm formation (Moormeier & Bayles, 2017). Biofilms have the ability to undergo series of nutrient recycling hence promoting the growth and survival of microorganism within the matrix (Bamford et al., 2023). Understanding the biology of biofilms makes it clear how important their complementary tactics are for both the microorganisms and the surrounding EPS matrix to either prevent the initiation of a biofilm or disrupt existing biofilms (Koo et al., 2018). Antibiotic penetration into biofilms rely mainly on the EPS structure which confers impermeability to large molecules of antibiotics. The outer EPS layer which resembles a capsule limit the entry of various antibacterial. According to the study by Mosaddad et al., 2019, the EPS matrix can expel the harmful molecules out of the matrix rather than allowing their entry into the biofilms. Extra-polymeric substances have the ability to impede the actions of antibiotic activities by means of enzymatic breakdown pathways and diffusion-reaction. Although, the composition of EPS differs throughout different biofilms, it often comprises of lipopolysaccharide and alginate, which work together as a barrier to the diffusion of antibacterial drug (Macià et al., 2014). Some biofilms contain residues of mannuronic acid and guluronic acid in their extracellular polymer shell which may operate as virulence factors to prolong infections by inhibiting immune responses and by shielding the biofilms from antibiotics, such as Ciprofloxacin, Gentamicin and Ceftazidime (Mirghani et al., 2022).

2.2.2 Developmental stages of biofilm formation

Biofilm developmental stages, as depicted in Figure 2.1, have been divided into four major, well-regulated events: (i) Initial attachment (ii) Biofilm multiplication (iii) Maturation (iv) Dispersal. An individual planktonic cell reversibly binds to a surface, and if the cells do not dissociate, they will bind irreversibly to the surface during the initial attachment. Surface proteins referred to as MSCRAMMs (Microbial Surface Components Recognizing Adhesive Matrix Molecules) facilitates this attachment (Foster et al., 2014). During infection, these proteins, such as fibrinogen, fibronectin, and collagen, play an important role in attachment to host factors (Lister & Horswill, 2014).

In the presence of a sufficient nutrient source, the adhered *S. aureus* cells will start to divide and accumulate following adsorption to a surface. Cell division and the production of the extracellular polymeric matrix are the two important processes that leads to biofilm maturation. Some proteins, like the SdrC, FnBPs, and ClfB proteins, play important roles in attachment and accumulation of biofilms. Serine-aspartate repeat-containing protein C (SdrC) at the cell-cell adhesion stage of biofilm formation engage in low-affinity homophilic bonds that promote intercellular adhesion (Foster et al., 2014). Fibronectin binding proteins (FnBPs) are also involved in cell adhesion stage allowing cells to bind together as the biofilm accumulates (T. J. Foster, 2016). The Clumping factor B (ClfB) promotes colonization of *S. aureus* in the host, facilitates biofilm formation, and causes virulence by binding soluble fibrinogen for immune escape (Abraham & Jefferson, 2012).

The final stage of biofilm development is seed dispersal, in which microcolonies separate in response to genetically programmed responses that mediate the seed dispersal process (Archer et al., 2011). Biofilm matrix dispersion can be mediated by proteases, nucleases, and proteins with surfactant activity. At this stage, some exo-polymeric substance components are broken down, allowing bacteria to escape from the biofilm (Kranjec et al., 2021). In this way, microcolonies migrate from the original site of infection to unaffected regions of the host system to enhance the continuous formation of biofilm and promote infection spread. Interest has been drawn towards biofilm dispersal as a means of treating chronic infections because dispersal aids the exposure and killing of metabolically active cells, making them vulnerable to the effects of antibiotics and the immune system (Kumar Shukla & Rao, 2013; Lauderdale et al., 2010). Additionally, dispersal mechanisms might be adapted to prevent the formation of biofilm on medical implants (Lazar et al., 2021; Opdensteinen et al., 2021).



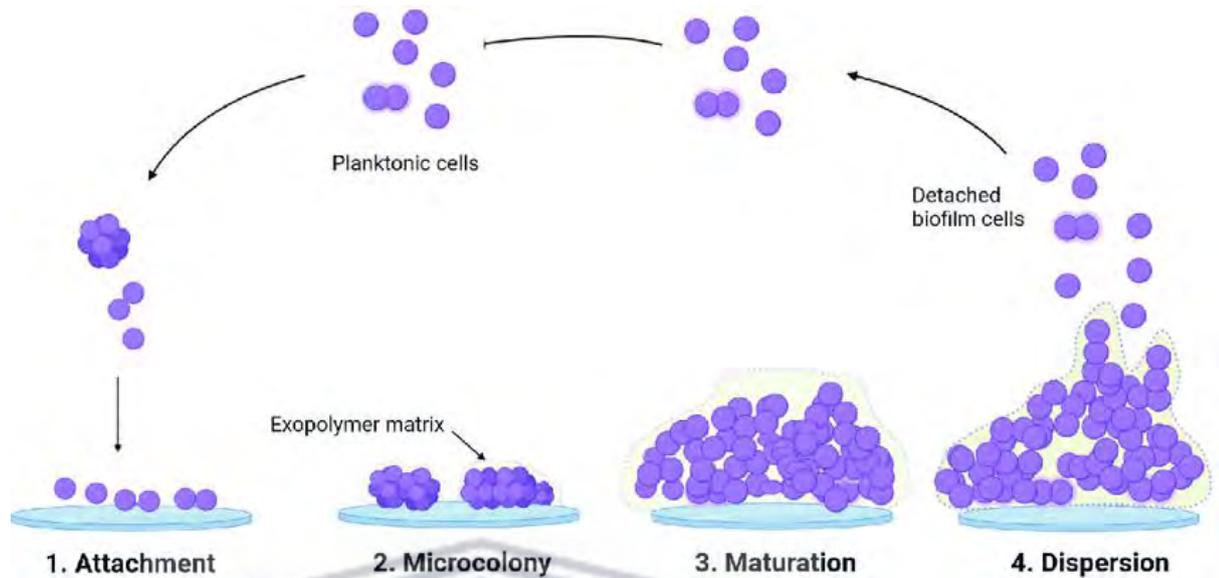


Figure 2.1 Developmental stages of *S. aureus* biofilm (Adapted from Alves Carneiro et al., 2020)

2.2.3 Clinical burden constituted by Staphylococcal biofilms

Biofilm-producing pathogens such as *S. aureus* have become well-known for causing persistent and chronic infections in humans. Conventional antibiotics are used to treat Staphylococcal biofilm-associated infections, but their misuse and overuse led to a 10 to 1000-fold increase in antibiotic resistance strains brought about by biofilm formation (Divakar et al., 2019; Steinig et al., 2019). Some antibiotics, like aminoglycosides, fluoroquinolones, β -lactams, and are inert against the bacteria in inner anaerobic biofilm because the antibiotics become inactive when oxygen and nutrient are not present (Mirghani et al., 2022). Bacterial biofilm is not only recalcitrant to the effect of antibiotics but also to the host immune system. Additionally, antibiotic resistance is also accelerated by poor infection prevention and control. Increased mortality and morbidity have

been associated with biofilm-related infections and infected medical implants, often requiring surgical treatments and prolonged hospitalization. As a result, there is an increased cost associated with the treatment of *S. aureus* infections (Cheung et al., 2021).

2.3 Epidemiology

If clinicians are to choose appropriate therapy, it is crucial to comprehend the prevalence of resistance in *S. aureus*. There is a distressing increase in antimicrobial resistance prevalence. Multidrug-resistance has prompted the licensing of new antimicrobial agents but resistance to the more recently introduced antibiotics has also emerged.

Staphylococcus aureus bacteria can be classified as both a commensal and a dangerous human pathogen. It is a major contributing factor to clinically significant infections including those that affect the skin, soft tissues, lungs, bones, contaminated prosthetics and medical devices (Tong et al., 2015). A few prospective studies have revealed a higher incidence of *S. aureus* infection in Africa than in industrialized countries. In South Africa, the annual incidence of *S. aureus* bacteremia was 3.28 cases per 1000 hospital admissions. Mozambique reported between 101–178 cases per 100 000, with the highest incidence in children under the age of five. In Kilifi, Kenya, the prevalence of SAB among children that are 5 years of age was 27 per 100,000 (Schaumburg et al., 2014). The study “A meta-analysis prevalence of resistance of *S. aureus* to different antibiotics in Nigeria” by Ezech et al., 2023, reported that the prevalence of resistance of *S. aureus* to different antibiotics ranges from 13 to 82%. Results showed a very high degree of resistance to Penicillin-G (82%), Cloxacillin

(77%), Amoxicillin (74%), Cefuroxime (69%), Ampicillin (68% [95%]). Moderately resistance to Erythromycin (47%), Chloramphenicol (47%), Methicillin (46%), ofloxacin (24%) and Rifampicin (24%). Low resistance was observed in Vancomycin (13%). The incidence of invasive MRSA in the black population (66.5 per 100,000 person-years) in the United States was reported to be more than twice that in the white population (27.7 per 100,000) person-years. This shows that there is a likelihood of a prevalence of SAB is associated with ethnicity (Jernigan et al., 2020). In Australia, the indigenous population have 5.8 to 20 times higher incidence of *S. aureus* bacteremia (SAB) than the non-indigenous Australians. Although socioeconomic status differences between indigenous compared to non-indigenous populations has a role to play, but does not fully explain the disparity between these groups (Hewagama et al., 2012; Tong et al., 2012).

2.3 Current approaches, success, and challenges to discovering antibiofilm compounds.

Once an antibiotic is clinically proven to be effective and is widely used for therapeutic purposes, its days are numbered because resistance that is clinically significant appears over periods of months or years. There is a struggle in the development of new approaches to combat infectious diseases due to the failure of existing antimicrobial therapies, which is of serious concern in the health community (Fleming & Rumbaugh, 2017). Several approaches, such as nanotechnology, quorum sensing, biofilm dispersal strategies, bacteriophages, and biosurfactants have been used in the discovery of antibiofilm compounds.

2.3.1 Nanotechnology

Nanotechnology provides a great platform for improving the physicochemical characteristics of various compounds to develop effective antimicrobials (Huh & Kwon, 2011). Ag nanoparticles enhance the antibacterial and antibiofilm activity in a synergistic manner. During iron oxide nanoparticle treatment, there has been a report of a substantial decrease in the growth of *S. aureus* and *Pseudomonas aeruginosa* biofilms on biomaterial and pluronic-coated surfaces (Thukkaram et al., 2014). Chitosan nanoparticles and ZnO-eugenol combination has been shown to effectively stop bacteria biofilm formation within the sealer–dentin interfaces of root segments (Dasilva et al., 2013). At higher concentrations of the Au nanoparticles, the formation of microbial biofilms was also inhibited (Sathyanarayanan et al., 2013). There is proof that nanomaterials could be used to prevent biofilm from forming on medical and biomedical equipment and food packaging materials. The toxicity of these nanoparticles to cells and biomolecules is of great concern, and they have restricted their clinical applicability (Joris et al., 2013). Currently, toxicological information about the effects of nanoparticles on human health is not readily available. However, new findings indicate that toxicity of multi-organ systems when antibiofilm nanoparticles are therapeutically administered by the formation of reactive oxygen species (ROS) because of interaction between nanoparticles and cell materials (Singh et al., 2017).

2.3.2 Quorum sensing

Targeting quorum sensing (QS) to regulate the virulence of bacteria can also be a tactic to control diseases (O’Loughlin et al., 2013). Because quorum sensing

plays an important role in microbial infections and biofilm development, finding compounds that are capable of interfering with QS in pathogenic bacteria is an emerging field for researchers. Three main types of quorum signalling exist: Gram (-) bacteria use N-acyl homoserine lactone (AHL)-based signalling, Gram (+) bacteria use autoinducing peptide (AIP)-based signalling and some autoinducer-2 (AI-2)-based signalling is found in some Gram (-) and Gram (+) bacteria (González-Ortiz et al., 2014). Different quorum sensing inhibitors have shown clinical advantages when used in combination with other antimicrobials. Structure-based virtual screening (SBVS) and *in-silico* docking analysis were used to look for potential quorum sensing inhibitors of *P. aeruginosa* (Lu et al., 2019). However, a limited success rate exists despite their applicability in clinical settings due to poor solubility, delivery, bioavailability, and stability. In many cases, studies frequently do not incorporate quorum sensing-independent controls, and toxicity is only determined by evaluating the impact on growth in a complex growth medium. Hence, as a result, the evidence for quorum sensing disruption is not always very strong, and many compounds that are presented to be quorum sensing inhibitors may actually be false positives when further studies are done (Defoirdt et al., 2013; Gorske & Blackwell, 2006).

2.3.3 Use of Biofilm dispersal strategy

The use of biofilm dispersing agents as a strategy has become an intense area of study because dispersed bacterial cells are typically more responsive to antimicrobial treatment than bacterial cells that are embedded in a biofilm matrix. Varieties of promising dispersal agents have been discovered (Fleming & Rumbaugh, 2017; Verderosa et al., 2019). A gel preparation composed of

Dispersin B and the disinfectant Triclosan has been sold for the treatment of skin and wound infections as well as the disinfection of medical equipment. The drawback is that; Dispersin B cannot be used to treat systemic biofilm-mediated infections because the bacterial enzymes have immunogenic properties. The use of dispersal agents can also be problematic despite being promising because if dispersal cells are left untreated, there is a possibility of translocating and seeding an infection in new sites, resulting in the spread of the initial infection. Therefore, dispersal agents are used concurrently with other antimicrobial agents to exert a synergistic clinical effect (Marvasi et al., 2014; Reffuveille et al., 2015). But, ensuring the concurrent presence of dispersal agents and antimicrobial agents in the correct concentration at the target site for treatment is challenging in the clinical setting (Fleming & Rumbaugh, 2018). Another challenge is the fact that drug co-administration of dispersal agents and antibiotics for treatments can lead to a higher risk of adverse effects, complex drug interactions, and side effects (Tamma et al., 2012).

2.3.4 Bacteriophages

Bacteriophages represent the most abundant biological entities on earth and constitute the absolute majority of life forms, and it is estimated that phage predation lessens the global bacterial population by half every 48 hours (Hendrix, 2002). The use of phage in therapy is increasing because of its recognised advantages over conventional agents. They are self-replicating and specific in action at the site of infection, thereby encouraging effective treatment of antibiotic-resistant pathogens. It involves the use of bacteria's natural predators; viruses that are capable of selectively infecting bacteria. Lytic phage interrupts

normal metabolism in bacteria and causes rapid lysis of bacteria (Carson et al., 2010). Bacteriophages alone or in combination with standard therapeutic agents are another attractive approach to consider in the treatment of biofilm infections. There are reports to show that bacteriophages are capable of degrading biofilm matrixes by initiating depolymerase production, which degrades the exopolymeric matrix components of the biofilm, leading to penetration of the inner layers of the biofilm (Azeredo & Sutherland, 2008). While some biofilms possess open structures with water-filled channels to give phage access to the inner biofilm layer, there is a diffusion limitation, particularly in dense biofilm structures. It is crucial to keep in mind that there are probably few chances of finding a specific phage with increased lytic capability and a specific host range. Genetic engineering of phages can be used to develop new genes with specific polysaccharide depolymerases, DNase, and proteases and to alter their host range to be able to efficiently degrade biofilms. The biofilm matrix serves as a storehouse of proteolytic enzymes as well as endoglucanases, which can inactivate bacteriophages. New genes can be engineered into the phages to enable them to destroy the bacteria found in the biofilm matrix. These isolated phages must be further characterized; an effective means of phage delivery and an in vivo analysis of phage performance need to be done to safeguard the phages against the human immune system (Azeredo & Sutherland, 2008).

2.3.5 Biosurfactants

Biosurfactants are another promising group of compounds that may be used in the treatment of biofilm-related infections. Biosurfactants and bioemulsifiers are used interchangeably in different kinds of literature. Because they have both a

hydrophobic moiety and a hydrophilic group, biosurfactants enable the presence of amphiphilic molecules at the interface between polar and nonpolar media. Biosurfactants (BS) prevents biofilm formation by reducing cell surface hydrophobicity, impeding the electron transport chain, distorting protein structure, interfering with quorum sensing, thus lowering the cellular energy demand (Satpute et al., 2016). Biosurfactants act by inhibiting biofilm physicochemical properties on the surface to reduce adhesion (Janek et al., 2012) and by downregulating bacterial gene expression involved in biofilm formation. The biosurfactant group is one type of chemical produced by microorganisms, and they are considered environmentally safe because they are biodegradable, biocompatible, and digestible. Different classes of biosurfactants are produced by different kinds of microorganisms possessing antibacterial, antifungal, and anti-biofilm activities (Paraszkiewicz et al., 2021). It was hypothesised that adding biosurfactant to mature biofilms causes rapid dispersion, changes the morphological changes of biofilm structures, and modifies the cell-surface hydrophobicity of the tested bacteria. This can ultimately impede the rate of deposition as well as the development of biofilm. Most of the biosurfactant compositions have not been fully investigated. Among the reports, about half (50%) of the 40 cell-associated biosurfactants omit structural details because their intricate structures are challenging to elucidate. Proteinaceous cell-associated biosurfactant and surlactin are the two that *Lactobacilli* spp. produce the most frequently (Satpute et al., 2016). Rhamnolipid (a biosurfactant)-silver and iron nanoparticle complexes have been demonstrated to be efficient against *Salmonella enteritidis*, *S. aureus*, *Bacillus pumilus*, *L. monocytogenes*, and

Yarrowia lipolytica (Paraszkiewicz et al., 2021). Although biosurfactants have enormous potential as antibiofilm agents, little is known about how toxic they are to humans.

2.4 Natural compounds as a potential source of anti-biofilm compounds

The need for agents that can prevent biofilm formation is urgent due to the important role that biofilm plays in infections and the emergence of multidrug-resistance (Lu et al., 2019). The existence of biofilm occurs in more than 90% of bacteria as a means of adaptive resistance to antimicrobial agents, which impedes effective treatment of acute and chronic infections (Li & Lee, 2017; Masák et al., 2014). This condition necessitates strategies to be developed for anti-biofilm agents that are specific and non-toxic.

Drug discovery and development are significantly influenced by natural products. For centuries, herbal treatments have been used in different cultures. There are recent reports that plant extracts are capable of regulating biofilm formation. For example, garlic extracts have compounds with antimicrobial activity, and quorum sensing is also inhibited by garlic extract (Bjarnsholt et al., 2005). Plant extracts from *C. trilobus* and *Coptis chinensis* have anti-adhesion effects at the adhesion stage of biofilm formation by inhibiting the membrane enzyme sortase (Kim et al., 2002). Rich in polyphenols, cranberry fruits influence the activity of proteolytic enzymes that inhibit the formation of essential elements of biofilms such as extracellular materials, carbohydrate production, proteolytic activities, and coaggregation (Duarte et al., 2006). *Pseudomonas aeruginosa* biofilm-associated genes were significantly inhibited by an extract from *Herba patriniae* (Fu et al., 2017). Phloretin, by the mechanism of efflux protein genes,

has anti-biofilm activity at low concentrations (1–256µg/ml) against *S. aureus* RN4220 and SA1199B (Lopes et al., 2017). Wheat-bran has shown the potential to obstruct bacterial quorum sensing systems by downregulating acyl-homoserine lactones (AHL), quorum-sensing signal molecules for gram-negative bacteria. The soluble extract of wheat bran at 5% showed anti-biofilm activity (González-Ortiz et al., 2014). Hence, natural products are regarded as a rich reservoir of bioactive compounds with therapeutic potential due to their remarkable chemical diversity.

Naturally occurring compounds are receiving immense attention due to increasing awareness about the side effects associated with the use of chemicals and traditional antibiotics (Chifiriuc et al., 2012). Up until now, a number of studies have examined the inhibitory effects of natural products on the formation and development of bacterial biofilm, suggesting their potential as a substitute for bacterial infection treatment (Lu et al., 2019). Natural compounds continue to be an abundant source of biologically active and diverse chemotypes. Natural therapeutic agents may have fewer side effects because they exert their physiological and pharmacological effects inside living cells. Natural products have a wider range of molecular properties, such as lower molecular mass, partition coefficient, and structural diversity although some natural compounds violate these. Additionally, natural products interact more with proteins, enzymes, and other biological molecules. Also, natural products have molecular rigidity and contain fewer heavy metals when compared with their synthetic alternatives (Mathur & Hoskins, 2017).

There are indications that more studies need to be performed because previous studies on plant-derived extracts failed to identify the molecular structures of antibiofilm bioactive molecules. For the development and assessment of natural antibiofilm agents in clinical applications, there is a need for enhanced efficacy and safety, either alone or in combination with other antimicrobial agents, which are essential to achieving great control of bacterial infectious disease in healthcare (Lu et al., 2019).

2.5 The use of computational tools in drug design

The use of high-throughput screening and combinatorial chemistry, which are traditional methods has led to an increase in the number of structural and biological data available to support rational decision-making in the pharmaceutical industries. This led to the introduction of a technique called cheminformatics (Gillet, 2019). Computational drug design is taking priority as a means of finding new therapeutics. Computational drug discovery is primarily used by chemist to reduce large compound databases into manageable sets of compounds with predicted activity that can be further tested experimentally. Also, through structure-activity relationship studies, it can be used to direct the optimization of binding affinity and pharmacokinetic parameters during lead compound optimization and to design novel chemotypes. It is not only aimed at explaining the molecular basis of therapeutic activity but also to predict possible derivatives that would improve the bioactivity of interest (Sliwoski et al., 2014). Structure-based and ligand-based computer-aided drug discovery are two subtypes of computational discovery.

Drug discovery processes can be made more effective by combining high-throughput screening and computational tools. Because predicted active compounds will be given priority and predicted inactive compounds will be skipped, this lowers the number of compounds for in-vitro screening while maintaining the likelihood of lead compound discovery, i.e., lowers the cost, time, and work required for high-throughput screening (Sliwoski et al., 2014). For instance, scientists at Pharmacia (now a division of Pfizer) effectively showed the potential of computational drug discovery when they used computational tools in parallel with HTS to screen 400,000 compounds for inhibitory bioactivity against tyrosine phosphatase-1B, an enzyme implicated in diabetes that hydrolyzes phosphotyrosines and inactivates insulin receptors. The outcome displayed two hit lists that were very dissimilar to one another. The docking hits were surprisingly found to be more drug-like than the HTS hits. The variety of both hit lists and how they differ from one another imply that docking and HTS may be complementary techniques (Doman et al., 2002). Examples of drugs that are discovered using computational tools include the Angiotensin-converting enzyme (ACE) inhibitor captopril (Talele *et al.*, 2010), carbonic anhydrase inhibitor dorzolamide (Vijayakrishnan 2009), saquinavir, ritonavir and indinavir, which are three drugs for the treatment of human immunodeficiency virus (HIV): (Drie, 2007), and tirofiban, a fibrinogen antagonist (Hartman et al., 1992). There was a recent study by Alves-Barroco et al. (2019) in which computational tools were used to screen molecules in the context of an antibiofilm agent to examine the presence of a biofilm regulatory protein BrpA homolog in *Streptococcus dysgalactiae subsp. dysgalactiae*

(SDSD) using high throughput virtual screening and molecular docking. In the study, five ligand molecules with high binding affinity to the hydrophobic cleft of the protein were chosen as potential inhibitor candidates for the SDSD BrpA-like protein.

2.5.1 Structure-based computer-aided drug discovery

Structure-based computer-aided drug discovery has gained widespread acceptance in drug discovery. Some libraries and databases are available for protein structures (Kalyaanamoorthy & Chen, 2011). Protein Data Bank (PDB) has the nuclear magnetic resonance (NMR) and crystallographic structure of proteins available for research purposes. Proteins are made up of amino acid residues, which are important for protein-receptor binding to ligands. Structure-based virtual screening is useful, especially when the 3D structure of the biological protein target is available. Structure-based computer-aided drug discovery requires prior knowledge of the protein's biological target to calculate binding energy and interactions for all ligands tested during screening.

The most widely used technique in structure-based virtual screening is molecular docking. Structure-based computer-aided drug design makes it simple to visualize the binding energy and binding mode of the ligands. Molecular docking adopts a scoring function for the different conformations to aid the analysis of the interaction between the protein and the ligands. The scoring functions are the mathematical methods put in place to predict the interaction between the protein and the ligands. To make scoring functions less complex, a lot of assumptions and simplifications must be made, which inevitably reduces their accuracy. The main flaw in docking in terms of accuracy is the existence of inadequate scoring

functions. The expensive calculations are still impractical for the analysis of numerous protein-ligand complexes and occasionally inaccurate (Kitchen et al., 2004). To solve this problem, using multiple scoring functions has been shown to improve accuracy and it leads to the identification of lead molecules. This is done by rescoring all the docked poses using a different function. Additionally, to evaluate and simulate the conformational space of a protein, additional computational methods like molecular dynamics and molecular mechanics are required (Sliwoski et al., 2014).

2.5.2 Ligand-based computer-aided drug discovery

Computational tools are available to make predictions from a set of compounds to prioritise them for further expensive in-vitro experimental studies to be able to identify compounds with desirable bioactivity (Gillet, 2019). This type of virtual screening adopted in studies depends on the type of available information. The goal is to retain the physicochemical properties most crucial for their desired interactions and ignore irrelevant data to the interactions. Ligand-based computer-aided drug discovery techniques are generally used when the target protein's 3D structure is unavailable and cannot be determined by homology modelling but information about active and inactive compounds of intended bioactivity is known (Sliwoski et al., 2014). For example, if the actives are known, a compound flexophore similarity search and pharmacophore mapping can be done to check other databases for compounds with similar activity (Khedkar et al., 2007). Also, if active and inactive compounds are known from bioassays, predictive classifier models can be built using machine learning tools to make predictions about compounds whose activity is unknown (An et al.,

2021). QSAR techniques can also be adopted in ligand-based computer-aided drug discovery (Neves et al., 2018).

2.5.3 Molecular similarity search

Molecular similarity search has long been adopted in cheminformatics. Three major components of molecular similarity are required: (i) a representation that encodes similar molecular and chemical features, (ii) a potential weighting of representational features, and (iii) a similarity coefficient (Gillet, 2019). The similarity coefficient ranges from “0” to “1”. The limitation of similarity search is that compounds sharing similar features can be identified ambiguously and are subjective except if they share the same pharmacophore. Similarity, like beauty, is more or less in the eye of the beholder. There are occasionally issues when attempting to quantify them and formally describe similarity relationships (Maggiore et al., 2014).

A small chemical modification can result in a significant change in bioactivity, a form of structural activity discontinuity (activity cliff) (Stumpfe et al., 2014). This can be explained with the concept of activity cliffs, which are groups of similar compounds and structures with significant activity differences. Activity cliffs may be traced to the presence or absence of specific receptor-ligand interactions, e.g., an important H-bond, a complementary fit of an aromatic group into a binding pocket, or an ionic interaction (Stumpfe & Bajorath, 2012). Similarity analysis has its place in drug development, but there is a need to further investigate these compounds obtained from similarity searches with other computational methods.

2.5.4 Machine learning

The popularity of machine-learning techniques is growing because of their capacity to make accurate predictions. In order to model, analyse, and predict various biological responses and processes during the drug discovery phase, pharmaceutical companies now frequently use machine learning tools (Lo et al., 2018). Machine learning algorithms are able to learn complex patterns from datasets to accurately forecast annotations on other data samples. A computer program uses a machine-learning algorithm to learn from experience (E) regarding the class of tasks (T) and performance (P) is measured, which improves with experience (E) (Tu, 2019). The input set of data with annotated labels (active and inactive) is called training data for the model to learn to make accurate predictions. To learn from the input data (compounds labelled as active and inactive), the machine-learning algorithm encodes a specific loss function. Loss function is the penalty that the learner incurs every time it commits an error, such as accidentally placing an active compound into an inactive bin during learning from the input training data. As a result, the algorithm gains the ability to learn to classify the input examples correctly. The learning algorithm is then capable of predicting the class label of the new compound (Zhang et al., 2021). Despite the potential advantages of machine learning in drug discovery, there are a number of obstacles that must be taken into account. The availability of appropriate data is one of the key obstacles in machine learning because machine learning-based approaches typically need a lot of data for training purposes. The accuracy and reliability of the results can be impacted by limited quantity, low quality, and inconsistent data. Modern machine learning-based approaches

cannot take the place of conventional experimental methods or the knowledge and expertise of human researchers. Machine learning can only make predictions based on the available data, and the results must then be validated and evaluated by human researchers (Blanco-González et al., 2023)

2.6 Conclusion

The National Institutes of Health claimed that biofilms are responsible for 80 percent of all chronic infections and 65% of all microbial infections, making biofilms a significant healthcare issue (Jamal et al., 2018). Biofilms are capable of settling on biological and non-biological surfaces, putting almost all patients at high risk, especially those with injuries, burns, inflammation, tissue damage, and patients with implanted devices and they can affect almost every organ of the body (Vestby et al., 2020). Planktonic bacteria have the ability to separate from a mature biofilm, spread to other organ systems, colonise them, and result in bacteremia or sepsis (Fleming & Rumbaugh, 2018). This is due to the fact that bacteria trapped in biofilms are extremely adaptable to antimicrobial treatment when compared to an identical bacterium in its free-floating planktonic state (Verderosa et al., 2019).

A global crisis such as the COVID-19 pandemic was evidently worsened by the overuse of antibiotics. Therefore, understanding biofilm formation and the development of antibiofilm agents to fight antibiotic resistance are priorities in the healthcare system (Strathdee et al., 2020). Despite this requirement and importance, there are presently no approved antibiofilm agents, and the majority of the previous studies used general antiseptics such as chlorhexidine or antibiotics such as cefazolin that are not biofilm-specific (An et al., 2021). The

difficulty in developing an antibiofilm agent is evidenced by the fact that there are no approved antibiofilm candidates despite ample years of research. Likely reasons for this could be the adoption of inaccurate models that show efficacy in in-vivo and in-vitro studies but fail in human studies, or because low priority is assigned to this class of drug (An et al., 2021). Adopting a combination of computational tools gives researchers the chance of discovering and developing some successful anti-biofilm agents. Computational methods that can be adopted to hasten the development of antibiofilm compounds against multidrug-resistant *S. aureus* have been reviewed in this chapter.



Chapter Three

Ligand similarity approach to discovery of potential antibiofilm hit compounds from two African natural product databases against multi-drug resistant *Staphylococcus aureus*

3.1 Introduction

Ligand-based drug discovery is an important aspect of computer-aided drug discovery. This method used in drug discovery is based on the principle that compounds that are structurally similar tend to have approximately similar biological activity. Important molecular fragment properties such as 2D properties, 3D properties, physicochemical properties, and flexophore descriptors are to be taken into consideration when looking for structural similarity because of their significant impact on the chemical characteristics necessary for binding to a target and to exhibit a desired pharmacological effect.

Molecular descriptors are mathematical functions applied to molecular representations to characterise molecular properties (Danishuddin & Khan, 2016). There are different types of descriptors, mainly: 1D, 2D, 3D, 4D, 5D, and 6D. 1D is based on the composition formula, 2D is based on the molecular graph, and 3D is based on the 3D molecular conformations. The 1D descriptor can explain the number of carbon atoms, number of heavy atoms and the number of carbon atoms, but it cannot explain related 2D or 3D information, for example, the number of aromatic rings, the number of double bonds, or the molecular surface area. Application of 1D and 2D topological descriptors has become

increasingly popular because these properties are derived from molecular structures using low computational resources. It is important to keep in mind that they do not encode conformational information, which limits their relevance for predicting the conformation-dependent properties of drugs (Helguera et al., 2008). An advantage of 3D descriptors is that they consider the 3D structures of ligands and are additionally applicable to sets of structurally diverse compounds. The major drawbacks of 3D-QSAR are that (i) it is not relevant to huge data sets containing more than several thousand compounds, which are usually taken into consideration in high-throughput screening, and (ii) 3D descriptors in QSAR analysis involve the computational complexity of conformer generation and structure alignments (Lo et al., 2018).

Comparing molecules is a frequent task in computer-aided drug discovery. Vector-based descriptors such as 2D and 3D have high performance for similarity calculations, but there could be a loss of information by condensing the molecular information into a descriptor vector. This is because the description of a molecule as a static arrangement of pharmacophore features cannot adequately describe its bioactivity. 4D, 5D, and 6D descriptors are multidimensional descriptors. They include the parameters involved in the structure and flexibility of the receptor-binding site in conjunction with ligand topology. 4D descriptors are based on reference grids and molecular dynamic simulations. Multiple conformations, orientations, protonation states, and isosteriomers are used to compute 5D descriptors, whereas solvation terms make up 6D descriptors (Peter et al., 2018).

A novel descriptor called flexophore fingerprint was introduced in the software Data Warrior (Sander et al., 2015). Flexophore compares similarities in chemical space orthogonal to chemical fingerprint descriptors of molecules while considering molecular flexibility (Von Korff et al., 2008). chemical space orthogonal to chemical fingerprint descriptors. Since the proposition of the induced fit theory, flexophore as a molecular descriptor has reflected dynamic conformational changes of molecules that occur during the recognition process of ligand binding to the target, which has been found to play a key role in drug discovery. Hence, molecules with similar flexophores are able to mimic similar bindings to the target of interest (Schuffenhauer et al., 2012). The descriptors similarly produced from the database molecules are compared to the query, and the database can then be sorted according to the similarity values of the flexophores. The comparison is based on the hypothesis that highly similar molecules to the query are more likely to be active than molecules that have a lower similarity i.e. the bioactivities of compounds that are structurally similar tend to be correlated more frequently than those of dissimilar ones (Cortés-Ciriano et al., 2020).

3.2 Method

Databases explored for this study are PubChem (<https://pubchem.ncbi.nlm.nih.gov/>), ChEMBL (<https://www.ebi.ac.uk/chembl/>), the aBiofilm database (<http://bioinfo.imtech.res.in/manojk/abiofilm/>), the South African Natural Compounds Database (SANCDDB; <https://sancdb.rubi.ru.ac.za/>) and AfroDb (Ntie-Kang et al., 2013) The data collection was done using a Dell PC Windows

10 with an Intel Core i5, a 64-bit operating system, and an x64-based processor. An electronic search was conducted in August 2021 on available databases.

3.2.1 Collection of active and inactive antibiofilm compounds against

Staphylococcus aureus

An electronic search was conducted on December 20, 2021 on the following databases PUBCHEM (<https://pubchem.ncbi.nlm.nih.gov/>), ChEMBL (<https://www.ebi.ac.uk/chembl/>) and aBiofilm database for reported active and inactive antibiofilm compounds against *Staphylococcus aureus*. The keywords used were “antibiofilm”, “*Staphylococcus aureus*”. The ‘aBiofilm’ resources (<http://bioinfo.imtech.res.in/manojk/abiofilm/>) harbour a database, a predictor, and the data visualisation modules. The database contains biological, chemical, and structural details of 5027 anti-biofilm agents (1720 unique) reported from 1988 to 2017. These agents target over 140 organisms, including Gram-negative, Gram-positive bacteria and fungi. They are mainly chemicals, peptides, phages, secondary metabolites, antibodies, nanoparticles, and extracts (Rajput et al., 2018).

From the literature studies, compounds with active and inactive antibiofilm activity of compounds against *S. aureus* were retrieved. Their respective bioassay ID, PUBCHEM CID, PUBCHEM activity, PUBCHEM standard value, PUBCHEM standard type, and PUBCHEM standard units were retrieved. According to the PUBCHEM assay presentation, compounds are marked as active if their activity is $\leq 50\mu\text{M}$ or if ChEMBL specifically reports them as active. From the aBiofilm database, compound ID, name, formula, and activity were retrieved. No literature reference standard was found for antibiofilm

percentage inhibition. Compounds with an inhibition percentage greater than 50% were marked as ACTIVE. For the activity expressed in folds, compounds with >2folds were marked as active.

3.2.2 Query databases

The query dataset of 411,180 compounds was collated from the South African Natural Compounds Database (SANCDb; <https://sancdb.rubi.ru.ac.za/>) and AfroDb (<https://doi.org/10.1371/journal.pone.0078085>), from which potential antibiofilm compounds against MDRSA will be identified in this study. SANCDb is a free database containing natural chemical compounds of South African origin. It was created in 2015 and has been useful in drug discovery studies for hit identification. On the other hand, AfroDb is a database of diverse natural compounds from African medicinal plants.

3.2.3 Retrieval of SMILES structures of datasets and generation of compound structures

The “Simplified Molecular Input Line Entry System (SMILES)” structures of active and inactive antibiofilm compounds and query dataset retrieved from SANCDb and AfroDb were used to generate their corresponding mol files of FragFp chemical structures using OSIRIS Datawarrior software (Sander et al., 2015).

3.2.4 Data characterization

To evaluate the structural differences between active and inactive antibiofilm compounds against *S. aureus*, the concepts of neighbour tree, scaffolds, flexophore similarity, and principal component analysis were adopted. The data

used for this chapter were the active and inactive compounds collected as previously described (3.3.1).

3.2.5 Calculation of compound properties and descriptors

DataWarrior software is capable of computing various properties of a compound directly from its chemical structure (Elaziz et al., 2018). In this study, Datawarrior was used to calculate and explore the chemical space diversity of collated active and inactive antibiofilm compounds, which is important for a comparative assessment of their structural and physicochemical properties. Flexophore descriptors of these compounds from their chemical structures were also generated. The key compound properties calculated were total average mol. weight in g/mol, cLogP, cLogS, H-acceptors, H-donors, relative polar surface area, topological polar surface area, electronegative atom count, stereocentre count, rotatable bond count, ring closure count, aromatic atom count, sp³-atom count, and symmetric atom count. Ring counts are also calculated, such as small ring counts, small ring count without hetero atoms, small ring counts with heteroatoms, small fully saturated ring counts, small non-aromatic ring counts, aromatic ring counts, small, saturated carbo-ring counts, carbo-aromatic ring counts, small carbo-non-aromatic ring counts, carbo-aromatic ring counts, small saturated hetero-ring counts, small hetero-non-aromatic ring counts, and hetero-aromatic ring counts. Box plots were generated for the molecular properties of active and inactive compounds with statistical significance set at $p < 0.05$.

3.2.6 Similarity charts and Scaffold analysis

This methodology of similarity charts is based on the principle that compounds that are similar would exhibit similar bioactivity. OSIRIS Datawarrior software (Sander et al., 2015) was used for similarity charts, and scaffold analysis. The similarity analysis calculates the entire fragment similarity between all compounds. The most similar neighbours are grouped together depending on their attractive forces, which increases with similarity. The resulting similar neighbour trees (clusters of compounds with similar features) are connected with a connecting line were analysed to identify the fragments that are related to the activity.

Scaffolds break down molecules into a framework of core structure and substituents. The scaffold analysis locates the core structure(s) of every molecule. The method used to locate the core structure(s) of the active and inactive compounds depends on the chosen Scaffold type. For this study, to generate the core structures, scaffolds for active and inactive compounds were analysed using the two different scaffold types: (i) Murcko scaffold: This includes every direct connection between the molecule's plain ring systems. Substituents, that do not contain ring systems are removed from rings and ring connecting chains and Ring systems with substitution patterns. (ii) Ring systems with substitution pattern: This mode identifies all annelated ring and single-ring systems without any substituents, but it also indicates that each ring atom has been substituted and that the original molecule contained an exo-cyclic, non-hydrogen substituent.

The scaffold frequency file option was selected for *DataWarrior* to create a new document listing all detected scaffolds and their occurrence frequency. Auto-SAR analysis was carried out using two scaffold types: the Murcko scaffold and the most central ring system. *DataWarrior* decomposes the structures by analysing scaffolds and substituents.

3.2.7 Flexophore similarity

The flexophore similarity concept is based on the claim that compounds with similar flexophore would have similar properties (Chhabra et al., 2021). A high flexophore similarity score indicates that the size, shape, and pharmacophore points of two molecules are comparable. The flexophore similarity scores were generated in OSIRIS DataWarrior software (Sander et al., 2015) based on the flexophore descriptor calculations of known active antibiofilm compounds to evaluate the query compound database with an automatic similarity limit of 85%.

3.3 Results and Discussion

3.3.1 Data collection and characterization

A total of 256 active compounds and 51 inactive compounds were collected from PubChem. A total of 67 active compounds and 54 inactive compounds were collected from aBiofilm database. Hence, there were a total of 323 active compounds and 105 inactive compounds (https://docs.google.com/spreadsheets/d/1XbScRvRuiK1-HBjmajK_9ro1N9Wfjz_s/edit?usp=drive_link&oid=116684118916762575224&rtpof=true&sd=true) . For the query database, a total of 411,180 natural compounds were compiled from both SANCDB and AfroDb

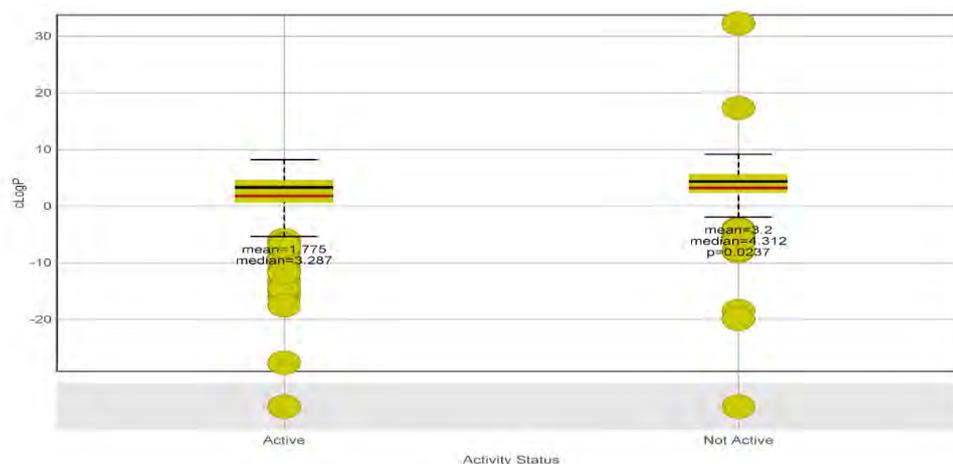


Figure 3.1(a) Box plot of cLogP of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively.

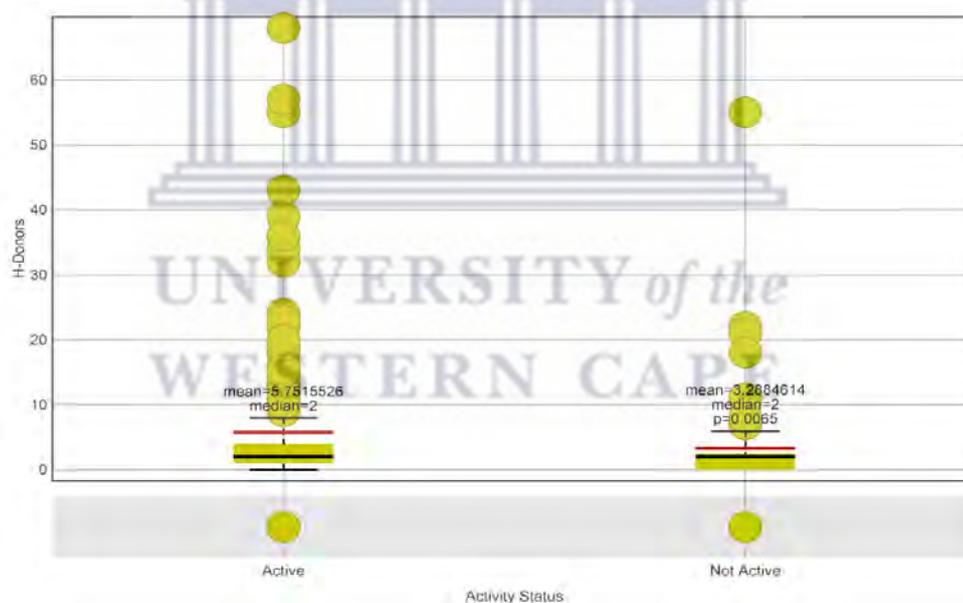


Figure 3.1(b) Box plot of H-Donors of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively.

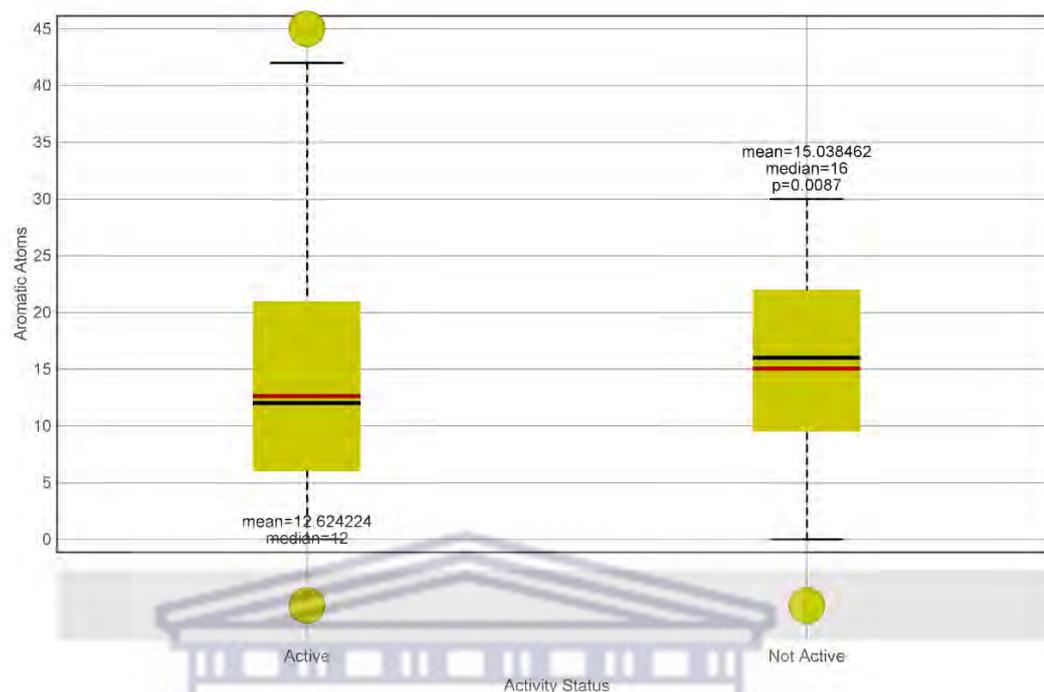


Figure 3.1(c) Box plot of Aromatic atoms of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively.

UNIVERSITY of the
WESTERN CAPE

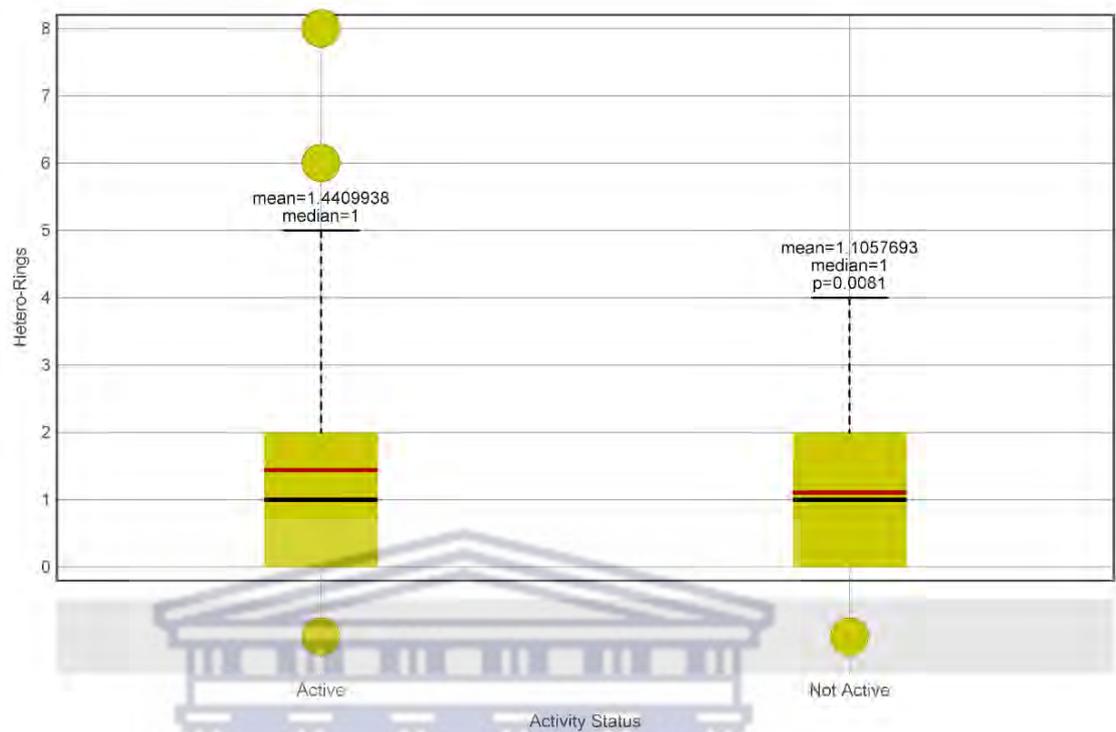


Figure 3.1(d) Box plot of Hetero-rings of active and inactive antibiofilm compounds against *S. aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively

UNIVERSITY of the
WESTERN CAPE

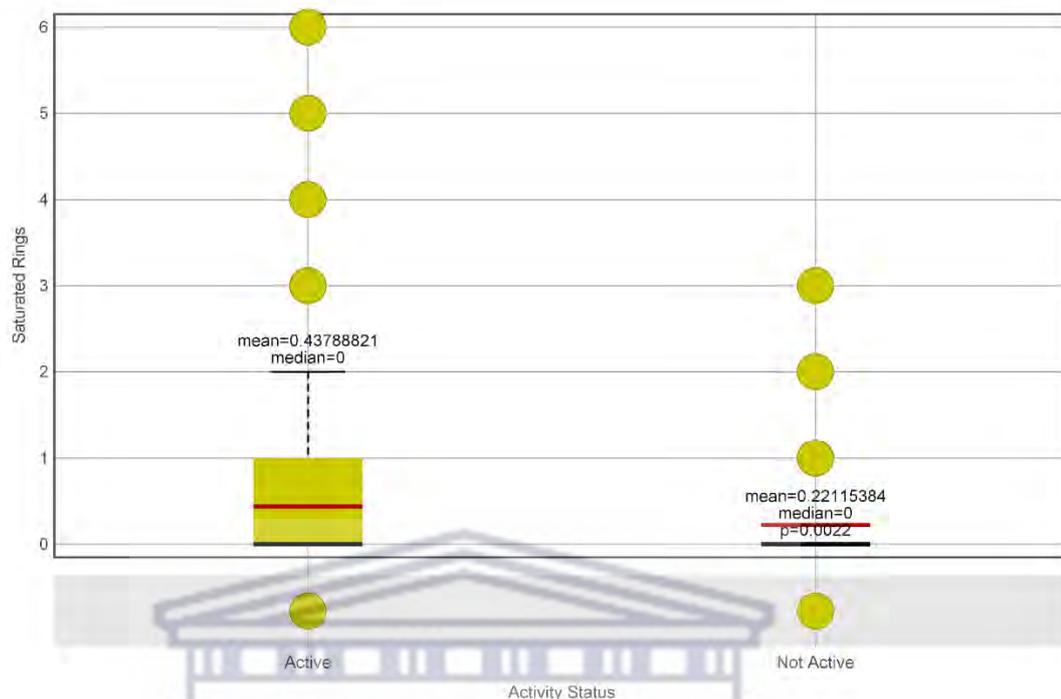


Figure 3.1(e) Box plot of Saturated Rings of active and inactive antibiofilm compounds against *Staphylococcus aureus* with significant P value <0.05. The red and black lines represent the statistical mean and median of each distribution respectively.

UNIVERSITY of the
WESTERN CAPE

The physicochemical properties of a compound could reveal the binding behaviours of the target of interest. The measure of hydrophilicity and hydrophobicity of a molecule expressed as a logarithm of the partition coefficient between n-octanol and water (cLogP) in this study shows that the reported active and inactive antibiofilm compounds have a cLogP value of <5. According to the Lipinski rule of 5, an oral drug should have a cLogP value of <5, and for good absorption, the ideal value is between 1.35 and 1.8. The active compounds in this study have a better cLogP value than the inactive compounds.

Compounds that are hydrogen bond donors can stabilise the 3D structure of binding sites. In general, H-bond donors are also less polar, thus enhancing the binding affinity at drug targets and also having an influence on the ADMET properties of a compound (Coimbra et al., 2020). This study reveals a higher number of H-bond donors among the active antibiofilm compounds. It was hypothesised that the more aromatic atoms present in a compound, the fewer the hydrophilic interactions and the higher the membrane permeability. This result shows that more aromatic atoms are seen in the active compounds than in the inactive compounds. A large number of heteroatoms in natural and synthetic compounds exhibit medicinal properties. There is no significant difference in terms of the presence of heteroatoms in the active and inactive antibiofilm compounds. It is observed that more saturated rings are present in the active compounds.

3.3.3 Similarity charts

An important concept in medicinal chemistry and drug discovery is chemical similarity, which notes similar compounds with enhanced bioactivities. There is

a well-established hypothesis that structurally similar molecules will have similar functions (Rao, 2021). The dataset was explored to see potential similarities in chemical structures between the reported active and inactive antibiofilm compounds using fragment and flexophore molecular fingerprints. The neighbour similarity trees are shown in Figures 3.2(a) and 3.2(b). Most similar neighbours are considered for every molecule; similar neighbours are connected by attractive forces. It was hypothesised that compounds that are connected in neighbouring trees will exhibit similar activity. In contradiction to this assumption, the result of the similarity tree shows compounds that are similar but have different activity profiles, i.e., some active compounds have cores (scaffolds) that are very similar to the ones present in inactive compounds. This could be explained by the concept of an activity cliff that manifests when a modest structural change significantly modifies the biological characteristics of the compound (Rao, 2021). To further get an insight into the structure differences observed in the similarity landscape, a close inspection of the structure in selected clusters (A, B, and C) was done, and the results are presented below in Figure 3.3(a-h). Analysing the resulting neighbour tree helps to determine the similarities and differences between the core fragments for both active and inactive compounds. The neighbour tree analysis shows no distinctive difference between the core fragments of known active and inactive antibiofilm compounds.

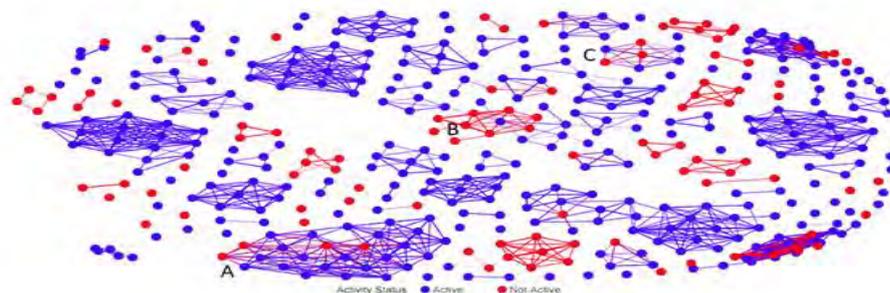


Figure 3.2(a) Resulting similarity tree for both active and inactive antibiofilm compounds using molecular fragment fingerprint. Active antibiofilm compounds denoted with red dots and non-active are blue dots. Similar neighbour compounds are connected with a connecting line to form a cluster.

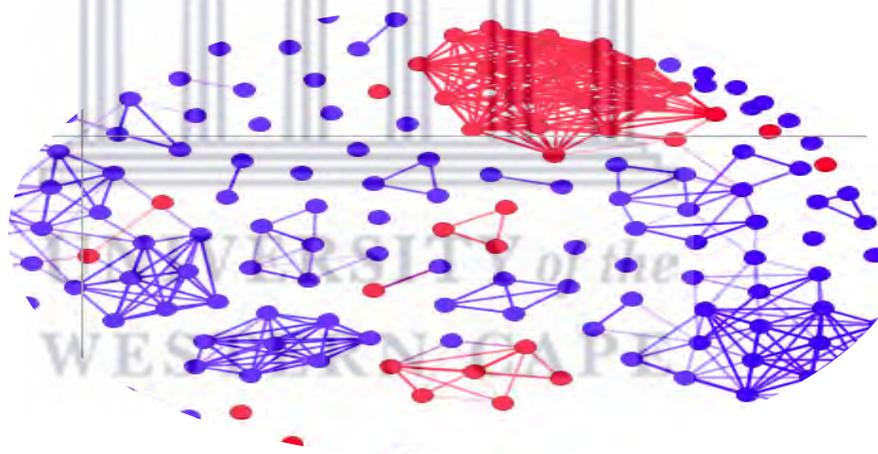


Figure 3.2(b) Resulting similarity tree for both active and inactive antibiofilm compounds using flexophore molecular fingerprint. Active and inactive antibiofilm compounds are denoted in red dots and blue dots respectively. Similar neighbour compounds with similar flexophore are connected with a connecting line to form a cluster.

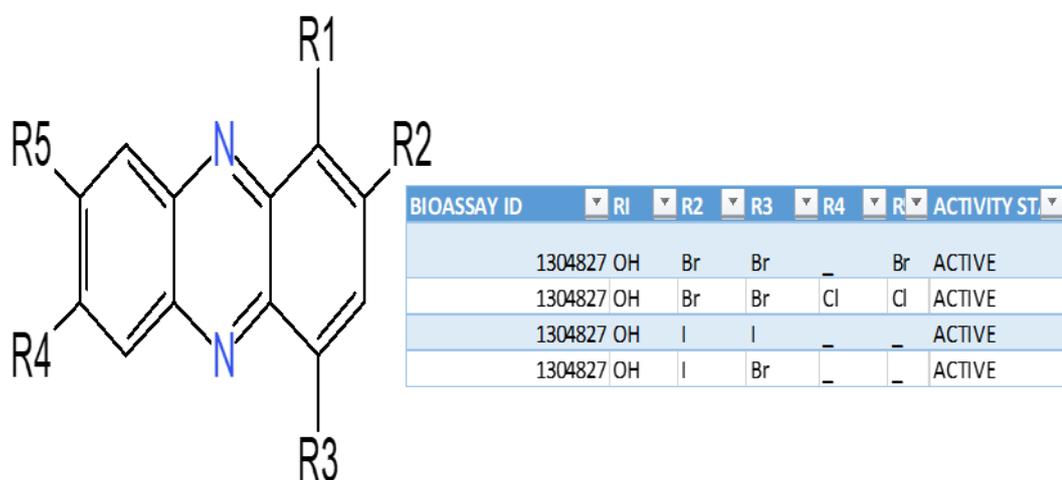


Figure 3.3(a) Analysis of neighbour tree A showing the core structure and substituents with their corresponding activity class

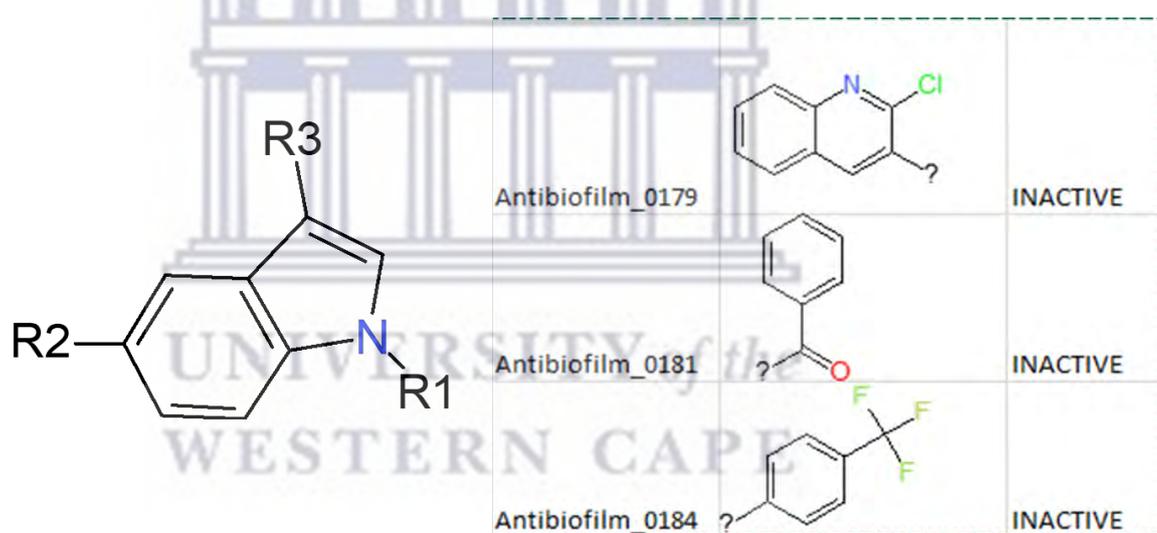
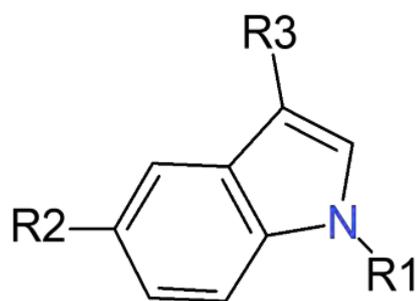
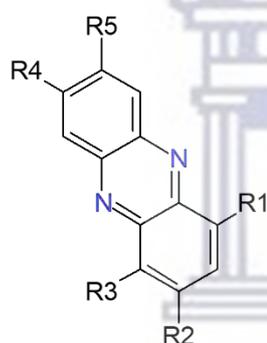


Figure 3.3(b) Analysis of neighbour tree A cont'd. showing the core structure and substituents with their corresponding activity class



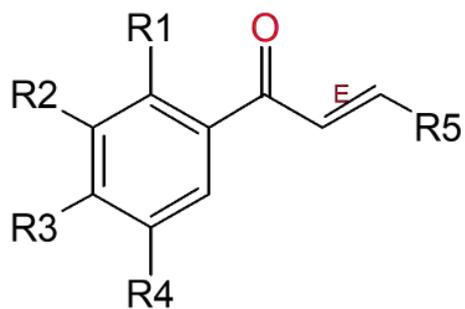
BIOASSAY ID	R1	R2	R3	ACTIVITY STATUS
1385102	H	Br		ACTIVE
1385101	H	Br		ACTIVE
1385102	?	Br		ACTIVE
1385102	?	Br		ACTIVE

Figure 3.3(c) Analysis of neighbour tree A cont'd showing the core structure and substituents with their corresponding activity class.



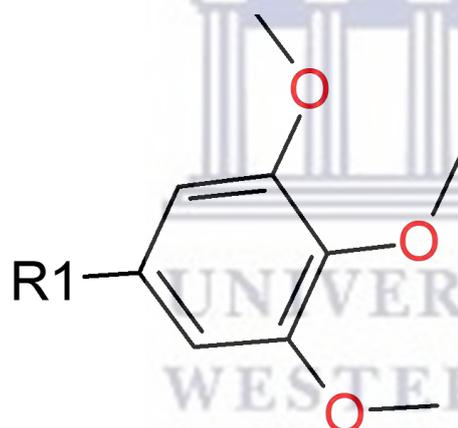
BIOASSAY ID	R1	R2	R3	R4	R5	ACTIVITY STATUS
1302847		Br	OH	-	-	ACTIVE
1302847		-	OH	Cl	Cl	ACTIVE
1302847	Br	Cl	OH	Br	Br	ACTIVE
1302847		Br	OH	Cl	Cl	ACTIVE

Figure 3.3(d) Analysis of neighbour tree A cont'd. showing the core structure and substituents with their corresponding activity class



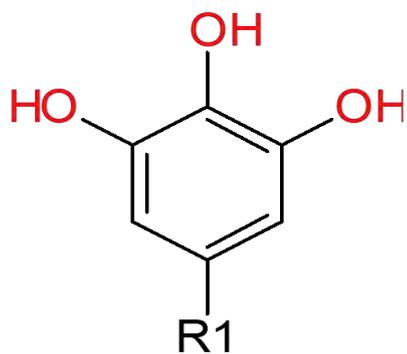
BIOASSAY ID	R1	R2	R3	R4	R5	ACTIVITY STATUS
Antibiofilm_0191		OCH3	OCH3			ACTIVE
Antibiofilm_0201	OCH3		OCH3	OCH3		INACTIVE
Antibiofilm_0203	OCH3		OCH3	OCH3		INACTIVE
Antibiofilm_0204	OCH3			OCH3		INACTIVE
Antibiofilm_0196		OCH3	OCH3			INACTIVE
Antibiofilm_0202	OCH3		OCH3	OCH3		INACTIVE

Figure 3.3(e) Analysis of neighbour tree B showing the core structure and substituents with their corresponding activity class



BIOASSAY ID	R1	ACTIVITY STATUS
Antibiofilm_0189		INACTIVE
Antibiofilm_0187		INACTIVE
Antibiofilm_0194		INACTIVE
Antibiofilm_0186		INACTIVE
Antibiofilm_0185		INACTIVE
Antibiofilm_0188		INACTIVE

Figure 3.3(f) Analysis of neighbour tree B cont'd. showing the core structure and substituents with their corresponding activity class



BIOASSAY ID	RI	ACTIVITY STATUS
Antibiofilm_0176		ACTIVE
Antibiofilm_0177		ACTIVE
Antibiofilm_0173		ACTIVE
Antibiofilm_0172		ACTIVE
Antibiofilm_0174		INACTIVE
Antibiofilm_0175		INACTIVE

Figure 3.3(g) Analysis of neighbour tree C showing the core structure and substituents with their corresponding activity class

UNIVERSITY of the
WESTERN CAPE

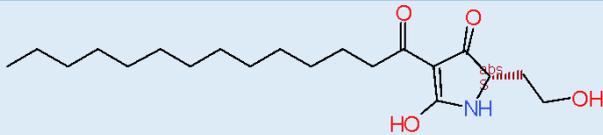
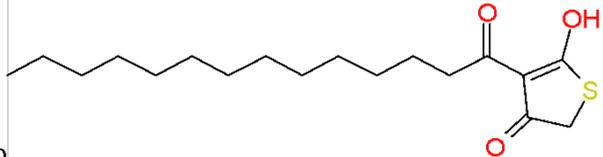
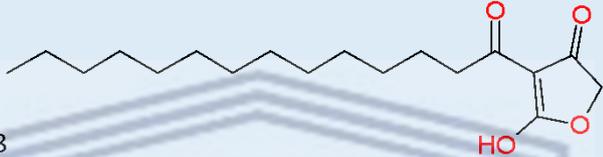
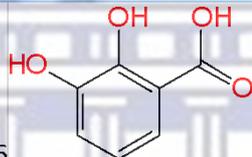
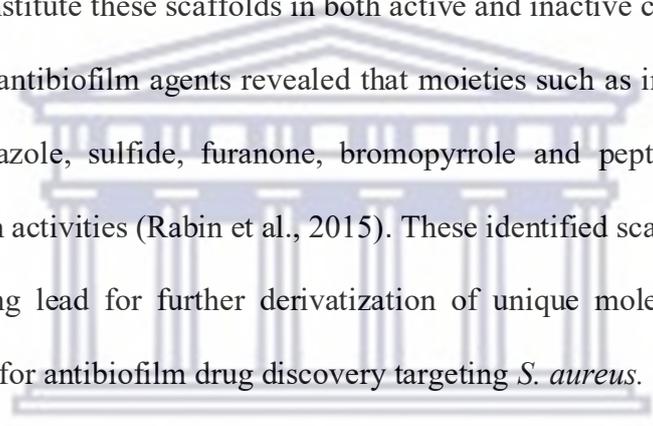
BIOASSAY ID	CORE STRUCTURE	ACTIVITY STATUS
Antibiofilm_3457		ACTIVE
Antibiofilm_3459		ACTIVE
Antibiofilm_3458		ACTIVE
Antibiofilm_1586		ACTIVE

Figure 3.3(h) Analysis of neighbour tree C cont'd. showing the core structure and substituents with their corresponding activity class

3.3.4 Analysing scaffold

There is a need to identify scaffolds in compounds with antibiofilm activity against *S. aureus*. The technique used to locate the core structure(s) depends on (i) the **Most central ring system**: in which the core structure is the molecule's ring system, that is closest to its topological center, and (ii) the **Murcko scaffold**, which has all plain ring systems of the given molecule and all direct links between them. Substituents that do not contain ring systems are removed from rings and ring-connecting chains. The scaffold diversity in terms of Murcko scaffold analysis demonstrates a thorough representation of diverse chemical scaffolds. Murcko scaffold analysis revealed unique scaffolds A, B, C, and D, as shown in figures 3.4(a) and 3.4(b), with varying degrees of frequency. Figure 3.4(c) listed

all identifiable scaffolds and provided a wider view of the functional differences between active and inactive compounds. There are scaffolds that are not outrightly similar between active and inactive compounds. Scaffolds A and D are present in both active and inactive groups but at a higher frequency in the active group. Scaffolds B and C in both the active and inactive groups are similar. Figure 3.4(c) compares the similarities between the scaffolds of active and inactive compounds. The pyrrole ring, indole ring, imidazole, and amide functional groups constitute these scaffolds in both active and inactive compounds. Recent studies of antibiofilm agents revealed that moieties such as imidazole, phenols, indole, triazole, sulfide, furanone, bromopyrrole and peptides contribute to antibiofilm activities (Rabin et al., 2015). These identified scaffolds can serve as a promising lead for further derivatization of unique molecules and diverse analogues for antibiofilm drug discovery targeting *S. aureus*.



UNIVERSITY of the
WESTERN CAPE

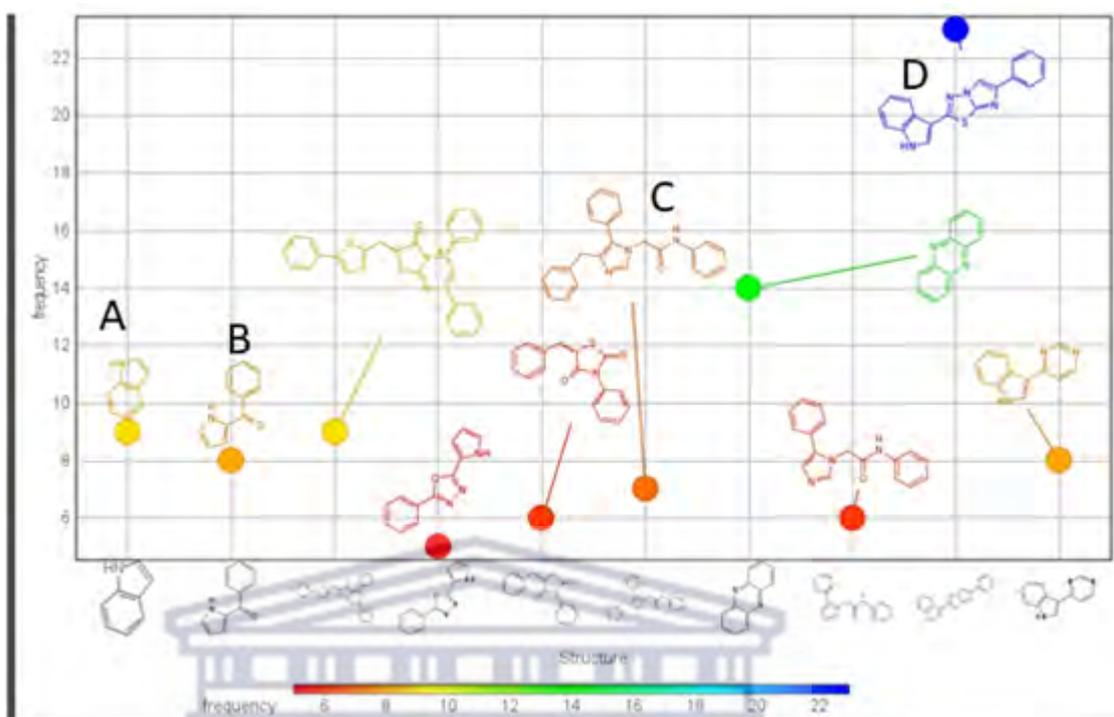


Figure 3.4(a) Murcko scaffolds of Active antibiofilm using Datawarrior. Unique scaffolds that are similar to those in the inactive group are labelled A, B, C, and D. The colours indicate the frequency of the scaffolds, where the blue and red colour represent smallest and largest values respectively.

UNIVERSITY of the
WESTERN CAPE

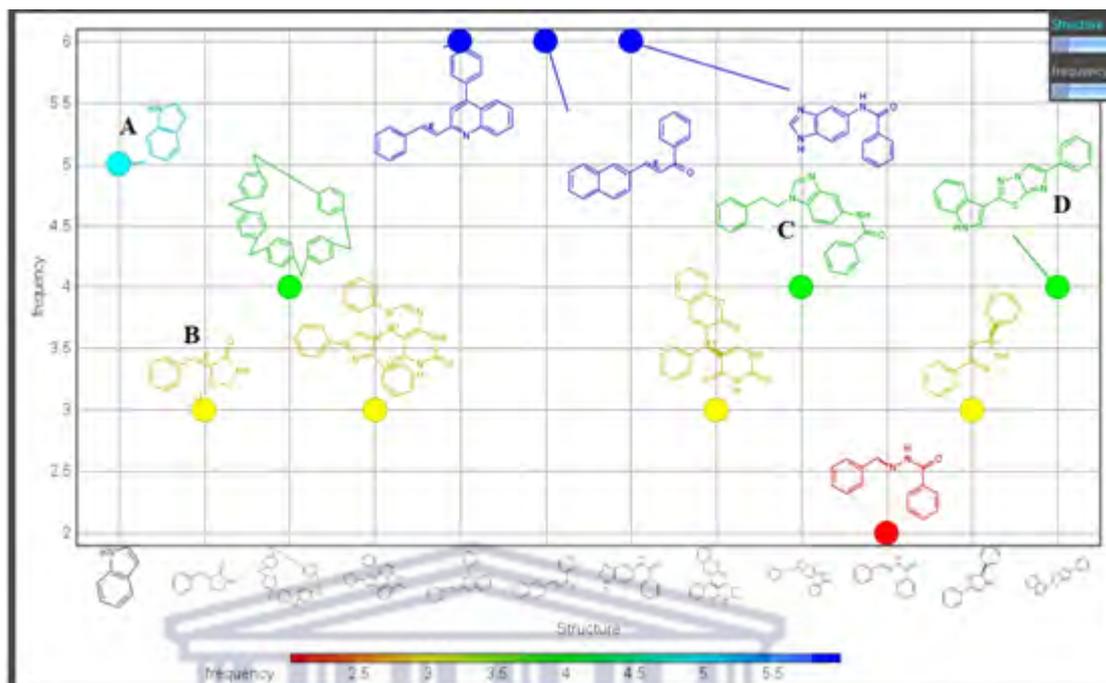


Figure 3.4(b) Murcko scaffolds of Inactive antibiofilm using Datawarrior. Unique scaffolds that are similar to those in the active group are labelled A, B, C, and D. The colours indicate the frequency of the scaffolds, where the blue and red colour represent smallest and largest values respectively.

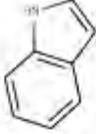
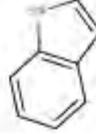
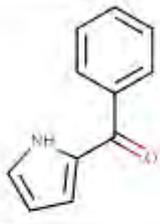
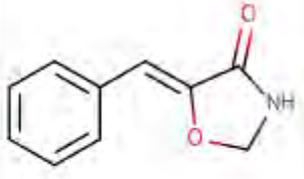
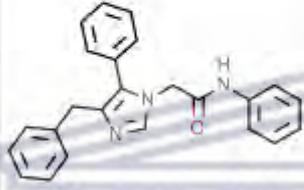
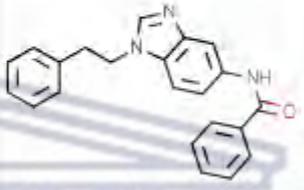
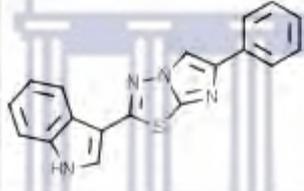
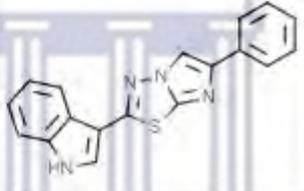
SCAFFOLD	ACTIVE	INACTIVE
A		
B		
C		
D		

Figure 3.4(c) Comparison between scaffolds A, B, C, and D of active and inactive antibiofilm compounds

3.3.5 Flexophore similarity search by comparing Active antibiofilm compounds and Query dataset.

The main approach is to compare the flexophore of compounds with known antibiofilm activity with the query dataset, whose antibiofilm bioactivity is not known. The selection of natural compounds in the query dataset was done using the flexophore similarity score to identify potential hits. The flexophore descriptor allows for predicting 3D-pharmacophore similarities. It provides a powerful and an easy-to-use way to see if any two molecules may have

compatible protein-binding behaviour. A high flexophore similarity signals that a significant portion of conformers of both molecules are compatible with regards to the size, shape, flexibility, and pharmacophore points (Sander et al., 2015). Compound flexophore similarity makes it possible to predict the biological behaviour of compounds because they are expected to exert their biological effects similarly, but they need to be subjected to further *in-silico* studies for validation to improve the reliability of the prediction. The flexophore descriptor calculation of active antibiofilm compounds was used to evaluate the query compound database. 43,957 compound pairs with flexophore similarity greater than 85% were generated from 411,180 compounds in the query database (<https://docs.google.com/spreadsheets/d/1maJToebF0QtmITYRB-3ldyqsHF4qpMDe/edit?usp=sharing&ouid=107704633229501699630&rtpof=true&sd=true>). This result demonstrated a significant level of flexophore similarity with the possibility that identified natural compounds will demonstrate similar antibiofilm bioactivity. The flexophore descriptors were able to enrich active molecules, where chemical similarity based on descriptors totally failed. Compound pairs of similar flexophore descriptors are shown in Figure 3.5 and Figure 3.6 presented a histogram showing the frequency distribution of flexophore similarity score. Overall, results from this study proves that the flexophore descriptors are capable of successfully encoding 3-D protein-binding behaviour rather than ligand FragFp chemical similarity. This is in line with studies from Von Korff et al., 2008 that showed that flexophore descriptors used to model biological similarity not only outperform chemical fingerprints but also

identify biologically active compounds where topological pharmacophore comparisons could not succeed.

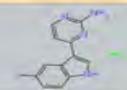
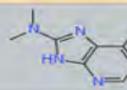
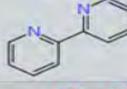
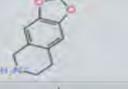
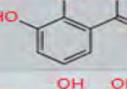
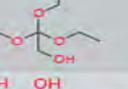
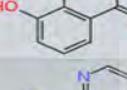
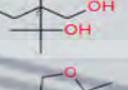
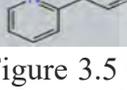
Structure 1	ID 1	Structure 2	ID 2	Similarity (Flexophore)
	1385102		CNP0285494	0.87138
	Anti-Biofilm_1620		CNP0210739	0.9435
	Anti-Biofilm_1586		CNP0293705	0.96271
	Anti-Biofilm_1586		CNP0313353	0.94471
	Anti-Biofilm_1620		CNP0018437	0.99264

Figure 3.5 Compound pairs of similar flexophore of reported active antibiofilm compounds (ID 1) to natural compounds in the query dataset (ID 2) with their corresponding chemical structure and flexophore similarity score.

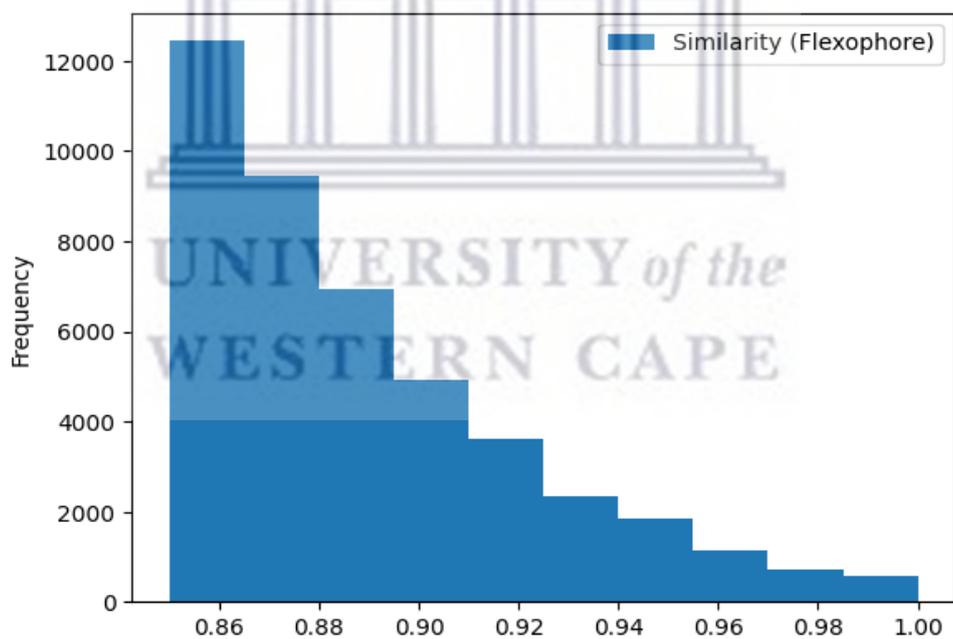


Figure 3.6 Histogram showing the frequency distribution of flexophore similarity score. Total count 43,957 of the query dataset, mean similarity score 0.88926, min 0.850000, Std 0.033, 25% 0.863030, 50% 0.880180, 75% 0.907270, max 1.000000.

3.4 Conclusion

Similarity charts/activity cliffs, scaffold analysis of active and inactive antibiofilm compounds resulted in no clear-cut difference and an overlap in functional groups identified. This occurrence can be explained by the concept of activity cliffs, in which structurally similar compounds have large differences in potency. Hence, flexophore as a molecular descriptor that expresses molecular flexibility became important for this study. The flexophore similarity metric highlighted the similarities between query datasets and the known active antibiofilm molecules, which would not have been detected simply by chemical similarity searches. 43,956 compound pairs with flexophore similarity greater than 85% were generated from 411,180 compounds in the query database. The pyrrole ring, indole ring, imidazole, and amide functional groups constitute the scaffolds in both active and inactive compounds. These identified scaffolds can serve as a promising lead for further derivatization of unique molecules and diverse analogues for antibiofilm drug discovery targeting *S. aureus*.

Chapter Four

Building a predictive model using a machine learning approach

4.1 Introduction

Staphylococcus aureus, with the emergence of antibiotic resistance has become an opportunistic pathogen that is capable of causing life-threatening infections (Dalman et al., 2019). Staphylococcal infection therapy is currently faced with many difficulties, not only due to the increasing resistance to the current antibacterial treatments and the multiple virulence factors it produces but also due to its biofilm formation ability (Jaśkiewicz et al., 2019). Machine learning as a branch of artificial intelligence, is becoming a promising pillar for overcoming the high failure rate in drug development. Machine learning methods have found extensive applications in predicting compound properties and in the area of drug discovery (Paraszkiewicz et al., 2021). Machine learning makes use of algorithms to analyse input training data, learn from it, and use it to make predictions on another set of related or unrelated data (Egieyeh et al., 2018). Currently, there is no antibiofilm predictive model for natural products with antibiofilm activity. Some publications are available for bioassays that report compounds with antibiofilm activity. In an attempt to discover potential antibiofilm compounds against multidrug resistant *S. aureus*, it may be expedient to learn from reported compounds from antibiofilm bioassays and thereafter predict the bioactivity of natural compounds from the query database. Machine learning can help in providing accurate predictions and ranking of compounds that could be further

tested for in-vitro and in-vivo activity. This will help reduce the extensive cost, time, and resources involved in antibiofilm laboratory bioassays. Hence, using machine learning approaches, reported antibiofilm bioactivity data of some compounds collated as described in Chapter three was utilised to build accurate antibiofilm predictive models. In this study, using a machine learning approach, four predictive models were built from the active and inactive antibiofilm bioactivity classes and a combination of molecular descriptors and molecular fingerprints in the dataset. The performances of the predictive models could be assessed with standard model evaluation parameters, namely: ROC (Receiver Operating Characteristic) area under the curve, and accuracy. The predictive models built in this study were used to screen for potential hit compounds from query dataset (SANCDb and AfroDb).

4.2 Materials and Methods

In this chapter, the aim is to evaluate the predictive power of models trained with datasets of compounds with reported antibiofilm activity against MDRSA and subsequently apply the trained model to the query natural compound dataset. Active and inactive compounds were compiled as described in Chapter Three. KNIME software (version 4.5.1) was used to construct and validate models that are capable of predicting antibiofilm bioactivity. An original Konstanz Information Miner (KNIME) workflow was set up as shown in Figure 3.1(a-d) for machine learning using active and inactive antibiofilm datasets so as to forecast the antibiofilm activity class of natural compound query databases.

4.2.1 Data

The dataset used in this study consists of the reported active and inactive antibiofilm compounds, collated according to the description in chapter three of this study. Konstanz Information Miner software (KNIME version 4.5.1) was used to construct and validate models that can predict antibiofilm bioactivity.

A total of 428 (75% active and 25% inactive) compounds were used in this study(https://docs.google.com/spreadsheets/d/1XbScRvRuiK1-HBjmajK_9ro1N9Wfjz_s/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true) . For the query database, a total of 411,180 natural compounds were compiled from both SANCDB and AfroDB (https://docs.google.com/spreadsheets/d/1ITW6n59hfLADsdjSqJTjLOhrQu_810-v/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true).

4.2.2 Machine learning algorithms

Four classifier algorithms were used to learn from the dataset: Multilayer perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF) and XGBOOST classifier. The specific classifiers were selected in order to represent four main types of classifier models: Random Forest represents tree-based classifiers; SMO represents function-based classifiers; XGBOOST is used in regression binary classifications and the Multilayer Perceptron represents neural network classifiers. The classifier algorithms were carried out with Weka (Waikato Environment for Knowledge Analysis) 3.6 nodes in KNIME.

4.2.3 Dataset pre-processing and calculation of molecular descriptors and molecular fingerprints

The input data was classified into active and inactive compounds. The machine learning experiment begins with data preparation, such as the generation of an RDKit molecule from Smiles, adding hydrogen to an RDKit molecule, and kekulizing an RDKit molecule. Fingerprints are generated for the molecules, and molecular descriptors are calculated for each molecule in the input table. The molecular descriptors were then normalised using a minimum-maximum normalisation node. The bit vector that represent the molecular fingerprint was expanded into individual columns for each compound.

4.2.4 Class Imbalance and cost-sensitive classification

The bioactivity class imbalance was identified as a major drawback to building an accurate model. As observed in this study, bioactivity classes in the datasets used are imbalanced because one class is overly represented (approximately 75% active class and 25% inactive class). Hence, the SMOTE (Synthetic Minority Over-Sampling Technique) node within KNIME was used to balance the bioactivity classes. To enrich the inactive instances in the training dataset, the SMOTE node oversamples the input dataset.

4.2.5 Selection of descriptors and features

The real-world data is noisy and may contain features that may be redundant, misleading, irrelevant, and that do not necessarily have good correlation with the output. The “backward feature elimination” loop was used in this study to filter

features to be applied during the building of models. The main idea behind feature selection is to select only features that shows strong correlation with the output. There are many kinds of feature selections methods — forward selection, recursive feature elimination, bidirectional elimination, and backward elimination. The simplest and the most widely used one is backward elimination. In forward feature selection, one feature is added at a time, and the addition is stopped when your model no longer improves or start to worsen. Backward feature elimination commences with a regression model that includes a full set of features and one feature is gradually removed at a time according to the feature whose removal makes the biggest improvement. The removal of features is stopped when the removal makes the predictive model to worsen. The core role of the “Backward Feature Elimination” meta-node in KNIME was to optimise the model prediction performance by feeding the model with the most significant descriptors and features that are crucial to building an effective classifier model. It is faster and more economical in building a predictive model (Saurabh Pal, 2021).

4.2.6 Training and Evaluation of performance of antibiofilm predictive models

Model training requires that an input dataset be split into a “training” dataset and a “test” or “validation” dataset (Nantasenamat et al., 2010). This process of gauging the model to the training data set is called “model training”. In this study, to learn from the dataset, a combination of molecular descriptors and fingerprints (circular FCFP6, circular ECFP4, MACC FCFP6, and MACC ECFP4) were

added to aid the accuracy of the predictions. Support Vector Machine (SVM), Multi-layer perceptron (MLP), Random Forest (RF), and XGBOOST were the four classifier learning algorithms used. To train the models, the input data were labelled as active and inactive which provided a baseline for each model to measure its performance against, helping them to learn the important features and descriptors over time and then analysing the relationships between the two classes of activity.

The “trained” machine learning models were validated, and their predictive performance was evaluated by using the test or validation dataset that was not part of the training dataset. The performances of all models were compared to determine the model with the highest accuracy. The models have prediction accuracy ranging from 0-1, where 1 indicates high biofilm inhibition activity. The performances of the classifier models were also evaluated by accuracy statistics and the receiver operating characteristic (ROC) curve after a 10-fold cross-validation of a training set and prediction of the bioactivity class of an independent test set. In the KNIME workflows (Fig.4.1), the scorer node and the ROC node were connected to the output from the predictor nodes. The area under the curve (AUC) value was also computed from the ROC curve. The results from the scorer node include the accuracy of the prediction and a confusion matrix. Accuracy shows the proximity of measurement results to the true value. The model with the best prediction accuracy was further used to evaluate the query compound database to predict their antibiofilm activity.

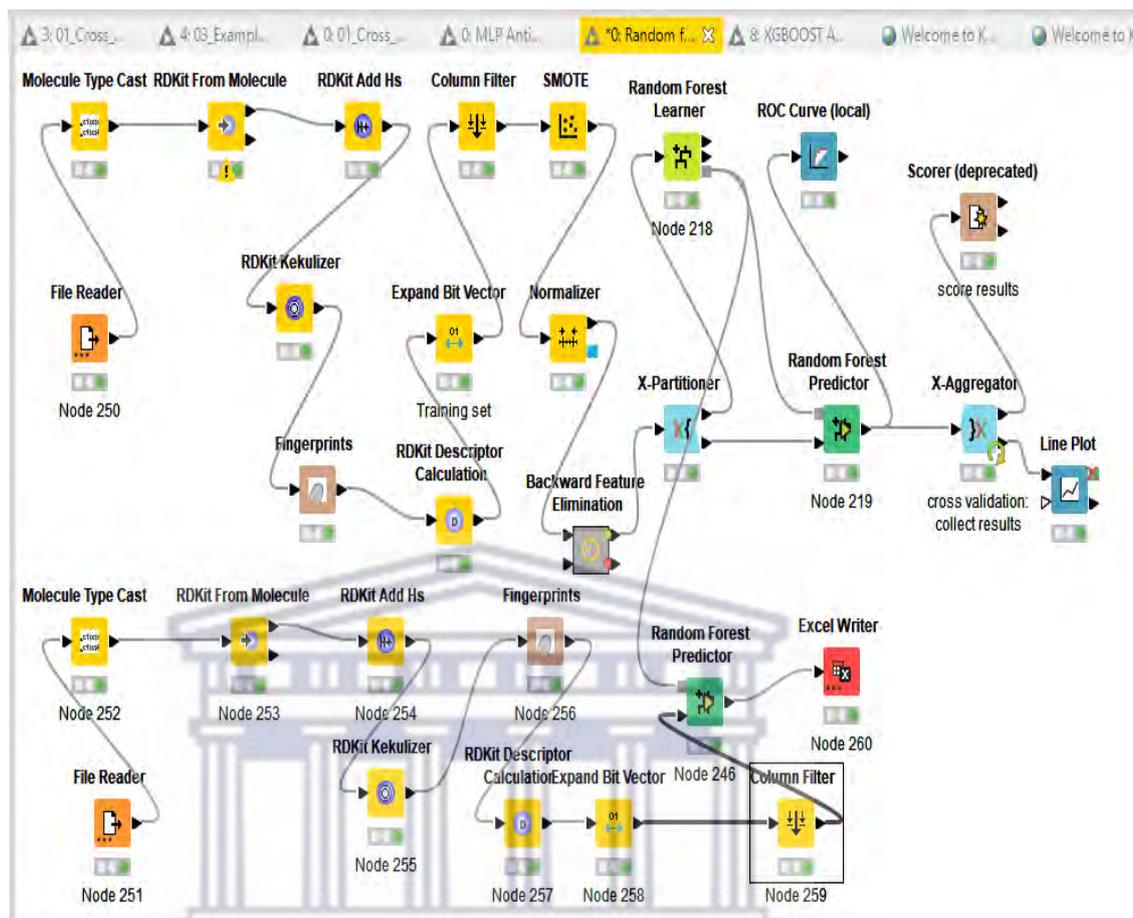


Figure 4.1(a) Screenshot of the KNIME workflow used to build the Random Forest classifier machine-learning model

UNIVERSITY of the
WESTERN CAPE

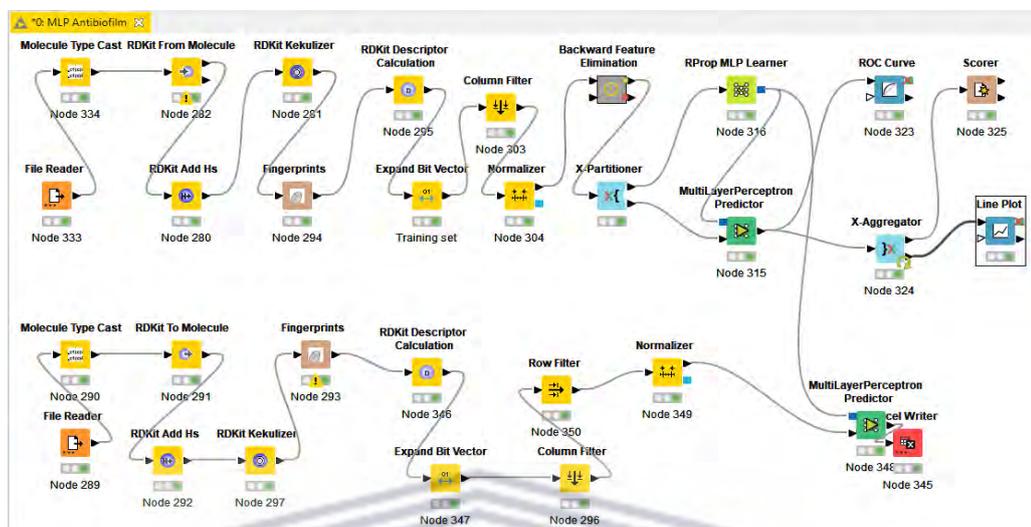
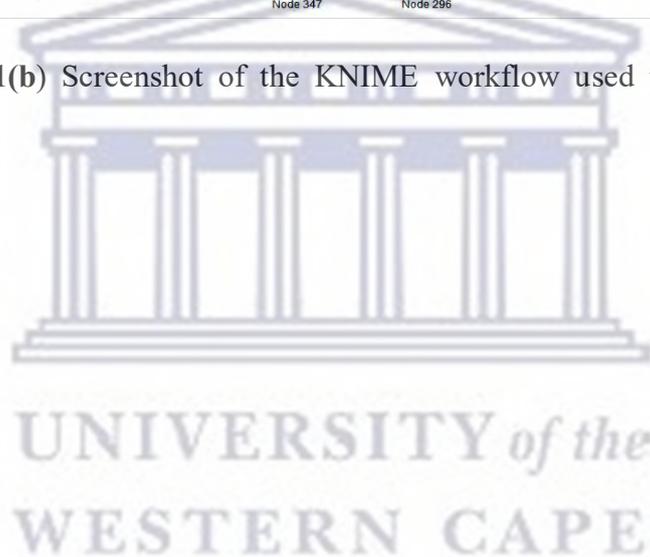


Figure 4.1(b) Screenshot of the KNIME workflow used to build the MLP classifier.



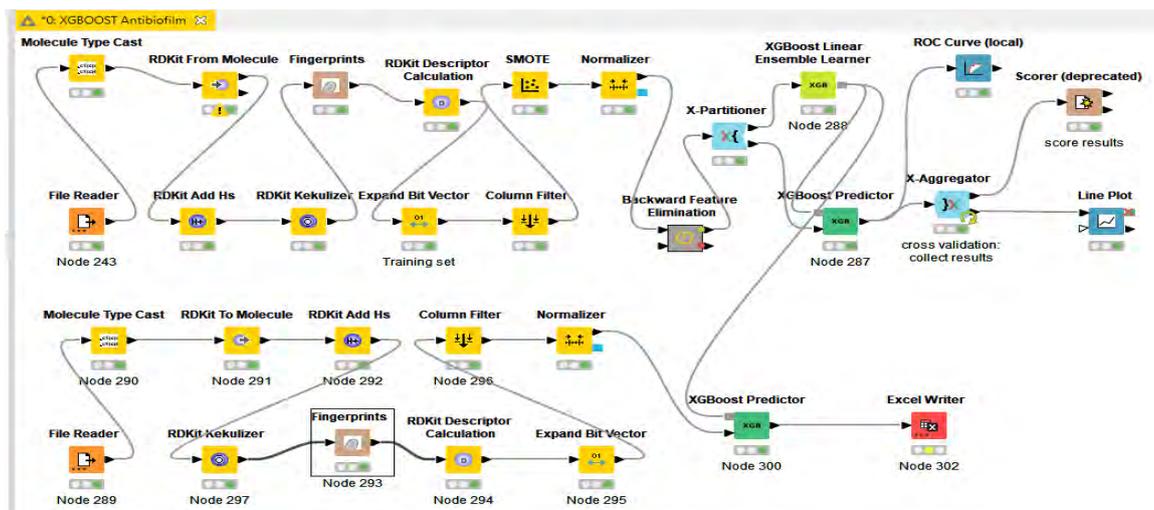


Figure 4.1(c) Screenshot of the KNIME workflow used to build the XGBOOST classifier machine-learning model.

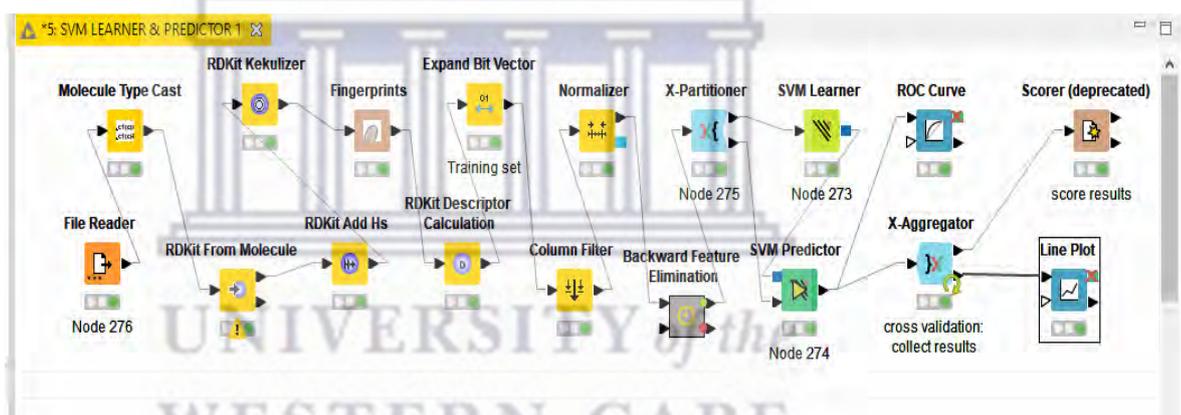


Figure 4.1(d) Screenshot of the KNIME workflow used to build the SVM classifier machine-learning model.

4.3 Results and Discussion

A total of 122 molecular descriptors were generated using the RDKit descriptor calculation node for the antibiofilm dataset: active = 325 (75%), inactive = 106 (25%). The resulting data was then passed on to the “Feature Elimination” metanode to remove redundant molecular descriptors combined with molecular fingerprints to train models. The class imbalance in bioactivity (75% active and

25% inactive) was identified as a limitation of these models because the active class is greater than the inactive class as represented in the dataset used. Hence, the “SMOTE” (Synthetic Minority Over-Sampling Techniques) node within the KNIME software was used to balance the bioactivity class. This node oversamples the input data by adding artificial rows to enrich the training data and adjust the class distribution. There could be different data division situations for training and test sets e.g., 70%–30%, 80%–20, and 90%–10%. “Training and testing data division influence on hybrid Machine Learning model process” conducted by Tao et al., (2020) revealed that 90%–10% data division attained better prediction capability. The dataset for this study was split into 90:10 to train and test the model. According to literatures, for a small dataset of less than 1000, 90%–10% data division is the best and to further improve the accuracy, 10-fold cross validation was adopted.

The values of the accuracy of the models is presented in Table 1 below. Accuracy is the proportion of compounds that were accurately classified as active and inactive after testing the trained model. The performances of the models were assessed by prediction accuracy statistics and the ROC (receivers operating characteristics) curve after cross-validation with an independent test set.

Molecular Descriptor+Fingerprint	SVM		Multi-Layer Perception		Random forest		XGBOOST	
	Accuracy	ROC curve	Accuracy	ROC curve	Accuracy	ROC curve	Accuracy%	ROC curve
Circular + FCFP6	75.82	0.649	86.85	0.825	95.97	0.995	93.79	0.977
Circular + ECFP4	76.76	0.578	87.97	0.892	95.96	0.986	92.70	0.995
MACC + FCFP6	84.74	0.915	87.09	0.897	93.48	1.000	93.17	1.000
MACC + ECFP6	84.51	0.783	88.26	0.944	94.26	1.000	92.24	1.000

Table 4 Value of the Accuracy and ROC curve of the models

From the results, the Random Forest (<https://docs.google.com/spreadsheets/d/17rcp47m-KJ4ZI-f85ZBAZt1-w6zVtQaW/edit?usp=sharing&ouid=107704633229501699630&rtpof=true&sd=true>) and XGBOOST model (https://docs.google.com/spreadsheets/d/1NTGa2zTwwJXmTrFlQT7yZXcN-ISZ9qgO/edit?usp=drive_link&ouid=107704633229501699630&rtpof=true&sd=true) showed similar results. These two models were further used to predict the antibiofilm activity of compounds in the query databases. XGBOOST model result is more difficult to interpret, it predicted over 90 percent of the compounds in the query dataset as active with varying degrees of prediction confidence that is as low as 0.5. Random forest model result for activity prediction on the other hand was easy to interpret. The compounds predicted are ranked according to the prediction confidence of the models. A total of 30,097 compounds out of 411,180 query dataset have Random Forest activity prediction of greater than 0.85. The machine learning model generated in this study helps generate hypotheses and hastens laboratory experimental verification in a less expensive and time-saving manner.

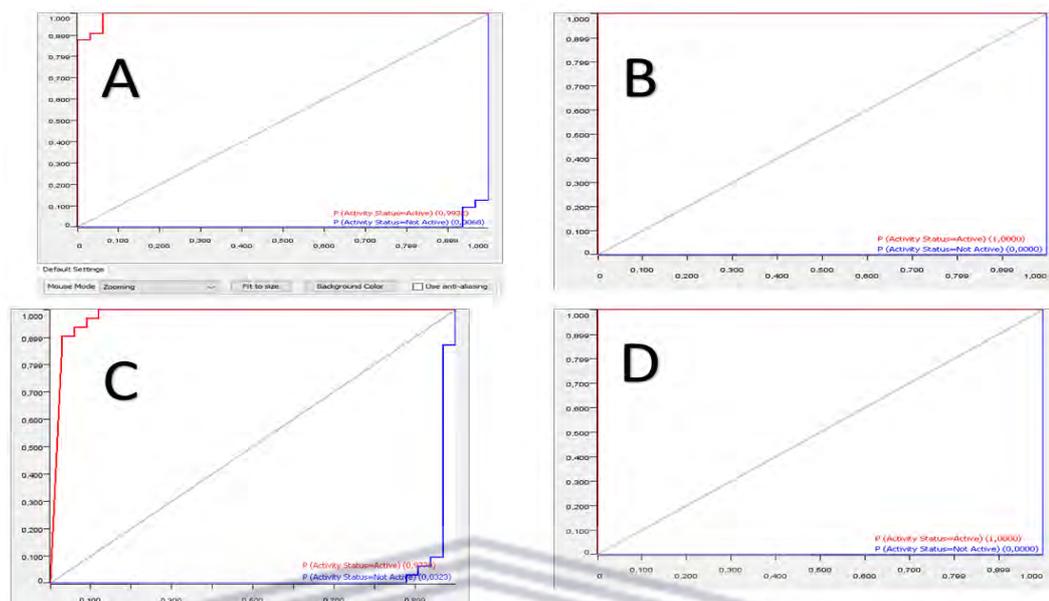


Figure 4.2 A and C represent ROC-AUC curve of Random forest and XGBOOST with B and D showing overfitting of the respective models with MACC+ FCFP6 fingerprint. Dark blue and red colour represents active and inactive compounds respectively. ROC-AUC plot evaluates model performance after training.

4.4 Conclusion

The machine learning approach was used to build antibiofilm predictive models that can predict the antibiofilm bioactivity of compounds. Important molecular descriptors and fingerprints of the active and inactive compounds combined were used for building predictive machine-learning models. Random Forest, XGBOOST, MLP, and SVM classifier models were built, but Random Forest and XGBOOST showed better predictive accuracy for the dataset. The Random Forest predictive model (accuracy 95.97%, ROC curve 0.977) was used to predict the antibiofilm bioactivity of natural compounds in the query database. Hit compounds with a prediction confidence of greater than 0.85 from the Random

Forest model were selected, which is a high threshold to avoid investing resources in compounds that are not promising for the drug development process. Overall, knowledge from this study could aid in the discovery of hit compounds that may be prioritised for the expensive processes of laboratory synthesis, in-vitro, and in-vivo bioactivity studies.



Chapter five

Consensus scoring for compounds from flexophore similarity search and Random Forest predicted model.

5.1 Introduction

The primary aim of this chapter is to use consensus scoring to rank the predictions of potential antibiofilm compounds from flexophore similarity studies and a Random Forest predictive model. This approach preferentially ranks the identified compounds to subject the highly ranked consensus-scored compound to further studies. A homologous set of initial scores is the prerequisite for this statistical consensus evaluation. In the early stages of drug discovery, scoring is often used to screen compound libraries for possible hits. The three commonly used statistical normalisation procedures are: (i) Ranking to represent docking scores for each target assigned against ascending ranks. This implies that ligands with more negative scores rank higher. (ii) Minimum–Maximum score scale (min–max scale) to rescale to a [0, 1] domain for each target and then deduct the score from the minimum score. The outcome is then divided by the difference between the maximum and minimum score. (iii) Z-score in which the min–max docking scores are mean-averaged or zero-centred and rescaled (Nhat et al., 2023).

5.2 Methods and Materials

Data for consensus scoring were the results generated from flexophore similarity search <https://docs.google.com/spreadsheets/d/1H-->

[9gd7Wr5VZoeQxAFTPvpC6Uz3g4eVa/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true](https://drive.google.com/spreadsheets/d/17rcp47m-KJ4ZI-f85ZBAZt1-w6zVtQaW/edit?usp=sharing&ouid=107704633229501699630&rtpof=true&sd=true)), and Random Forest predictive model (<https://docs.google.com/spreadsheets/d/17rcp47m-KJ4ZI-f85ZBAZt1-w6zVtQaW/edit?usp=sharing&ouid=107704633229501699630&rtpof=true&sd=true>). Top compounds from flexophore similarity search and Random Forest predictive model were compared to identify if they both predict the same set of compounds and also to identify compounds that are equally predicted as top compounds by both approaches.

The average mean score was employed as a measure of central tendency, i.e., a typical representative value of prediction. Min-max normalization technique subtracts the data values with the minimum and divides it by the range, i.e., the difference between maximum and minimum.

$$X^* = [X - \min(X)] / \text{range}(X)$$

$$X^* = [X - \min(X)] / [\max(X) - \min(X)]$$

where $\min(X)$ is the minimum; $\max(X)$ is the maximum; and $\text{range}(X)$ is the difference between maximum and minimum. The range is in the interval of [0, 1], and the length of the interval is 1 (Sinsomboonthong, 2022)

The z-score can be calculated by subtracting the population mean from the raw score, or data point in question and then dividing the difference by the standard deviation:

$$z = (x - u) / \sigma$$

where x is the score in question, u is the mean score, and σ is the standard deviation. (Id et al., 2018)

Histogram plots for the distribution of flexophore similarity score, random forest machine learning prediction confidence, Z-score normalized average, and the average mean score was done, where average score = (flexophore similarity score + random forest machine learning prediction confidence)/2.

5.3 Results and Discussion

Comparison of results of flexophore similarity search and Random forest predictive model shows that they both predicted two different sets as to compounds

(https://docs.google.com/spreadsheets/d/1a7joGY7B2Hsc8zbN39xa2fEIfW_ae61t/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true).

Top 30,097 compounds (approximately 10% of 411,180 query compounds) have flexophore similarity score >0.85

(https://docs.google.com/spreadsheets/d/1VQrJHuCvxepZAhJhPXSINwSeY2152OEp/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true).

A total of 45,576 compounds (approximately 7% out of 411,180 query compounds) have Random Forest activity prediction confidence of >0.85. To avoid taking false positive results or missing out important compounds that may show activity when subjected to further studies, the concept of consensus scoring of top compounds predicted by both approaches was adopted.

The main benefit of consensus scoring over individual virtual screening is its ability to reduce false positives and negatives (Nhat et al., 2023). Figure 5.1(a), (b), (c), and (d) shows the histogram plot for the distribution of Random Forest Machine Learning predictive confidence (RF-ML), flexophore similarity score, average score, and Z-score normalized average score respectively. The peak bars

in the histogram represent the most common values. In statistics, a measurement that describes the relationship of a value to the mean of a group of values is referred to as the Z-score. The Z-score is a measure of the standard deviation from the mean. If a Z-score is 0, it implies that the data point score is identical to the mean score. If the data point is above average, a positive Z-score will be observed. A negative z-score indicates the data point is below average. A Z-score close to 0 means the data point is close to average. A data point can be considered unusual if its Z-score is above 3 or below -3. From the Z-score result (https://docs.google.com/spreadsheets/d/1-rkUw_9b4819Dj4c6EYpMxOFvu3SPc2i/edit?usp=sharing&oid=107704633229501699630&rtpof=true&sd=true) and average mean score (https://docs.google.com/spreadsheets/d/1xYqMppaDW1CWflF6c-5jb5GKnMpatvLs/edit?usp=drive_link&oid=116684118916762575224&rtpof=true&sd=true), 99.9% of the z-scores are close to the average mean prediction score. For this study, a high threshold of 0.85 average mean prediction score was employed so that resources will not be invested in compounds that may later fail in the process of drug development.

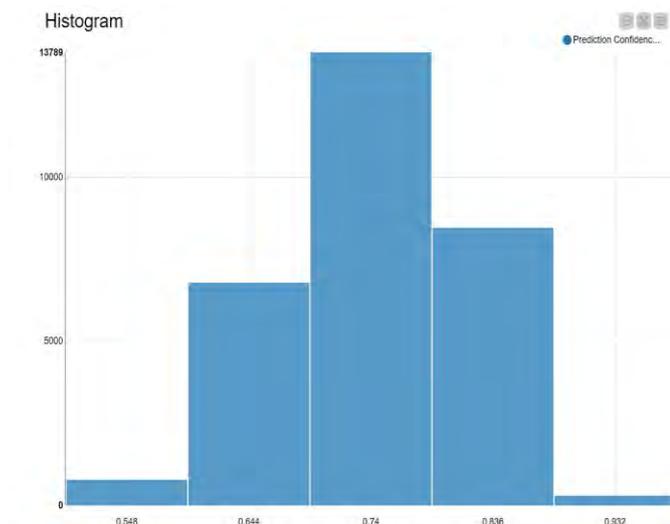


Figure 5.1(a) Showing the Distribution of RF-ML prediction confidence. The prediction scaled from ‘0’ to ‘1’, the higher the prediction confidence value, the better the chance antibiofilm property.

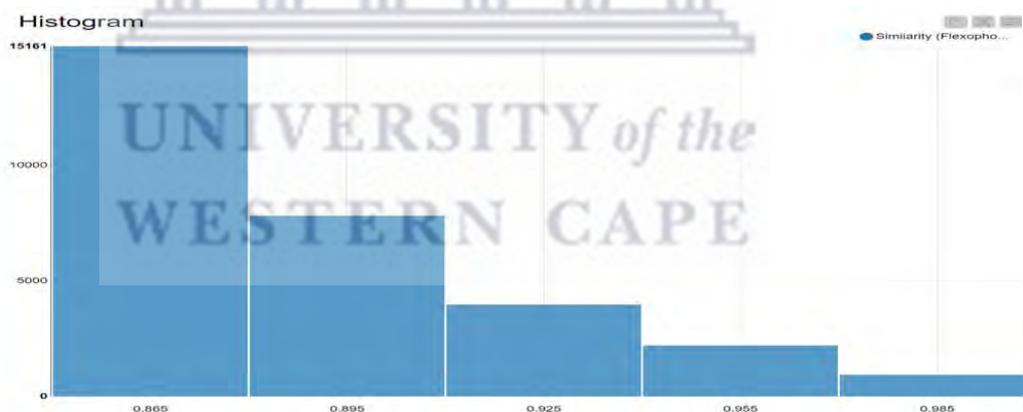


Figure 5.1(b) Histogram showing the distribution of flexophore similarity score. The score ranges from ‘0.85’ to ‘1’, the higher the score the greater the similarity between the natural compound flexophore and the known active antibiofilm compound flexophore

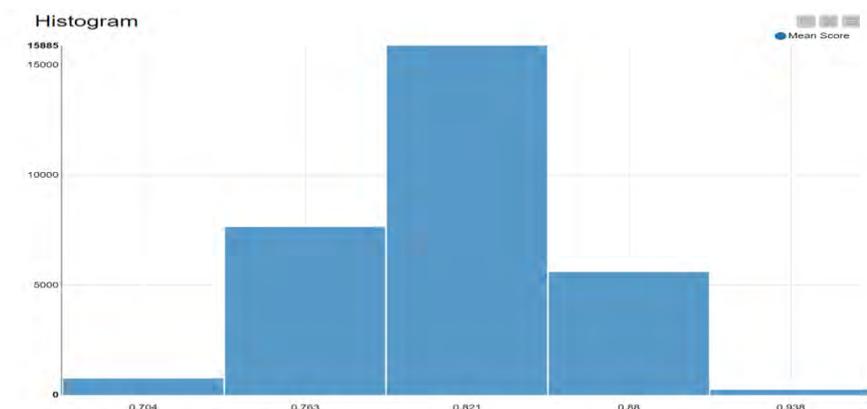


Figure 5.1(c) Histogram showing the distribution of average score = (flexophore similarity score + RF ML prediction confidence)/2

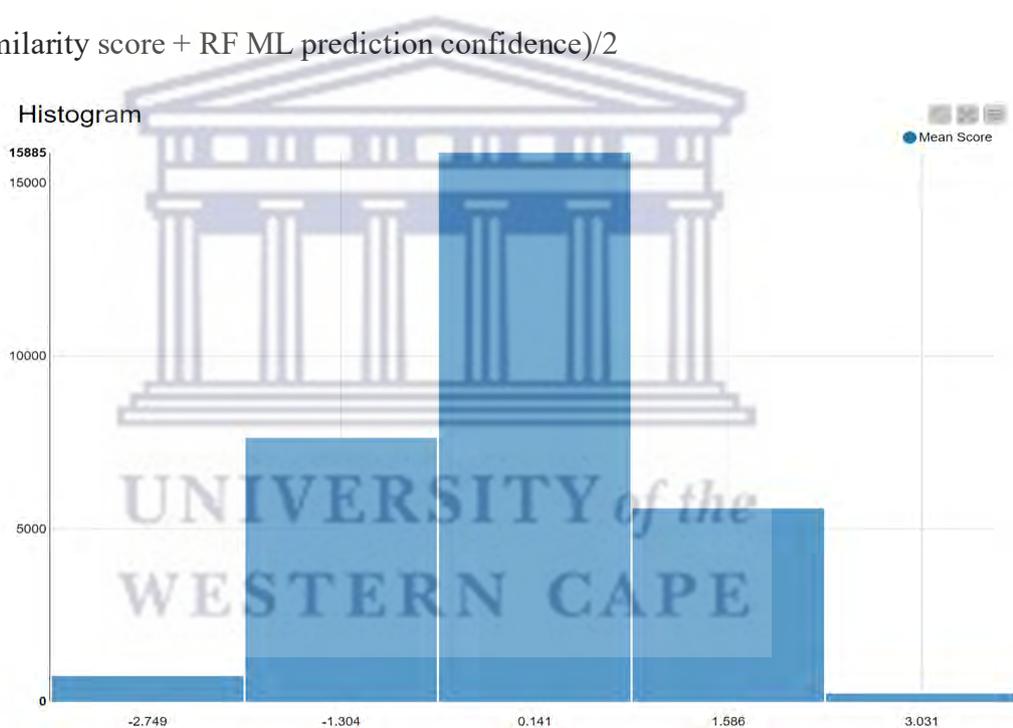


Figure 5.1(d) Histogram showing the distribution of Z-score normalized average score.

5.4 Conclusion

Consensus scoring techniques can be more robust and effective than using the broader set of available scores for the antibiofilm compounds from the ligand similarity searches and Random Forest machine learning approaches. Consensus

scoring combines different scores to compensate for errors from individual scoring functions, therefore improving the probability of finding the antibiofilm hit compounds.



Chapter Six

Reverse molecular docking of top ranking consensus-scored compounds with antibiofilm activity

6.1 Introduction

In the early stages of drug discovery, docking and scoring are often used to screen compound libraries for the identification of possible hits against a given protein target. A powerful computational tool for the identification of potential interactions between ligands and biological targets is molecular docking. The ability of one or more compounds to bind to a target protein can be evaluated *in silico* using reverse docking (RD). This strategy is useful for the identification of molecular targets of bioactive compounds, finding alternative uses for drugs, proposing new molecular mechanisms, or predicting drug toxicity (Chang et al., 2021). The aim of this chapter is to utilize reverse docking (RD) to understand the protein-ligand interaction and predict the possible mechanism of action for the antibiofilm activities of the top 142 identified compounds from mean and Z-score consensus scoring against multidrug-resistant *Staphylococcus* (<https://docs.google.com/spreadsheets/d/14YWvXRp-2bFSfLtsHqQL9sXQ0MdDm0VD/edit?usp=sharing&oid=116684118916762575224&rtpof=true&sd=true>).

6.2 Methods and Materials

Glide docking was performed using Schrodinger Maestro Release 2021-2. The *Staphylococcus aureus* biofilm-associated proteins for this docking study were

retrieved from the BioSIM database (https://drive.google.com/file/d/1CCAXfEbJgOYp0y8slGiQO7FvYTZ0aw1N/view?usp=drive_link).

6.2.1 Ligand preparation

Different tautomeric and protonation states are generated for the 142 ligands. A total of 225 ligand states were generated as Maestro output. In addition to generating states, Epik also assigns an energetic penalty for each ligand state.

6.2.2 Protein selection

A total of 39 *S. aureus* biofilm-associated proteins were compiled from the Biosim database (https://drive.google.com/drive/folders/1qyDMZC3giLuoVLuyFR9R7QgKQEFTdOUw?usp=drive_link). A significant database of protein sequences and their related in-depth annotation is the UniProt (<https://www.uniprot.org/>). Information about protein name, category, function, and subcellular location were retrieved from UniProt. 3TIQ, 5DBL, 4WVE, and 3TIP are SasG (*S. aureus* surface protein G) proteins that are involved in the adhesion stage of biofilm formation. Adhesion is a process in which planktonically growing microorganisms of identical species aggregate and develop on solid substrates while moving through a liquid and produce extracellular polymers that make attachment and matrix formation easier, which alters the organism's growth rate and gene transcription. SasG's fibrillary structure explains its ability to mask the binding of *S. aureus* microbial surface components by recognising adhesive matrix molecules (MSCRAMMs) to their ligands and promoting the formation of biofilm. These proteins are located sub-cellularly in the cell wall.

4AE5 trap protein, a transcriptional regulator located in the cell membrane is involved in signal transduction. The activation of the aggregation system and subsequent RNAIII synthesis phosphorylated TRAP results in the production of several virulence factors. It regulates the expression of the majority of toxins and genes known to be essential for the formation of biofilm. 3GEU (IcaR protein) is also a transcriptional regulator that represses transcription of the IcaADBC operon necessary for biofilm production. 7DM0, 7C7R, and 7C7U are matrix proteins involved in cell-abiotic substrate adhesion, cell-cell adhesion, and single-species submerged biofilm formation, i.e., the adhesion of a cell to an underlying abiotic substrate and the attachment of one cell to another cell via adhesion molecules. The subcellular location of these proteins is in the cell wall.

4B60, 4B5Z, 2RL0, 2RKZ, 3CAL, and 2RKY are fibronectin-binding proteins located in the cell wall. They have several interchangeable fibronectin (Fn) binding sites, each capable of promoting adhesion to both soluble and immobilised forms of Fn. This confers on *S. aureus* the ability to penetrate endothelial cells both *in-vivo* and *in-vitro* without the need for extra factors, although in a slow and inefficient way through rearrangements in host cells. This invasion process is facilitated by integrin alpha-5 and beta-1 promotes bacterial attachment to both soluble and immobilised forms of fibrinogen using a unique binding site located within the 17C-terminal residues of the gamma-chain of human fibrinogen. Both plasma proteins function as a bridge between the bacterium and host cell, promote attachment to immobilised elastin peptides in a dose-dependent and saturable manner, promote attachment to both full-length

and segments of immobilised human tropoelastin at multiple sites in a dose-dependent and pH-dependent manner, and promote adherence to and aggregation of activated platelets independently of other *S. aureus* surface molecules.

3AU0, 3ASW, 3AT0, 4F24, 4F20, 4F1Z, 4F27, 5JQ6 and 2VR3 are ClfB (clumping factor B, fibrinogen-binding protein B) located in the cell wall. They are cell surface-associated proteins that are linked to virulence because they promote bacterial adhesion to both alpha- and beta-chains of human fibrinogen, inducing the formation of bacterial clumps. 3BS1, 4XYQ, 4XXE, 4XQQ, 4XQN, 4XQJ, 4XY0, 4G4K, and 4BX1 are transcriptional regulators that are necessary for high-level post-exponential phase expression of a few secreted proteins. 3BS1 and 4XYQ are found in the cytoplasm. 7VF0, 7VFK, 7VFL, 7VFN, 7VFM, and 7EC1 are transferases. They belong to glycosyltransferase, a group 1 family protein. They are involved in the catalysis of the transfer of a glycosyl group from one compound to another.

6.2.3 Identifying binding sites in protein targets

The site map program on Schrodinger Maestro software (version 2021_2) was used to identify druggable binding pockets and rank the potential binding sites of the biofilm-associated proteins. Top-ranked sites were generated with their corresponding x, y, and z coordinates. The first-ranked site was selected for use in protein grid generation for docking. A druggable binding pocket is characterised by a favourable hydrophobic and hydrophilic balance. According to Michel et al., (2019), a druggability score (D-Score) of 0.80 is identified as a difficult target. D-scores higher than 1.1 are considered excellent drug targets. D-

score values smaller than 0.8 are undruggable. Undruggable targets are strongly hydrophilic, have no hydrophobic character, are relatively small in size, are very shallow, and require covalent bonding. Difficult targets are sufficiently hydrophilic and less hydrophobic, and they require administration as a prodrug that is cleaved in-vivo to produce ionic functionality that may be essential for ligand binding. They are classified as difficult targets because their design as prodrugs complicates the developmental process of drug design. Druggable targets are of reasonable size, and hydrophobicity with unexceptional hydrophilicity (Halgren, 2009). Druggability prediction is very significant because it allows focusing mainly on promising targets with good prospects for drug development. 17 proteins out of the 39 biofilm-associated proteins have a good druggability score of greater than 0.8 from the site-map

(https://docs.google.com/spreadsheets/d/10ShssbijX7kq-xK2QnwVMNQSSZMwBAip/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true).

6.2.4 Grid generation and Docking

Grid generation of the protein receptors is used to define the binding pockets. The binding sites of the proteins are buried at the centre of the grid box because glide docking does not deal with the entire protein structure but the grid. An inner box is set up at the centroid of the predicted binding site on the protein (Ban et al., 2018). GLIDE docking was used because it searches for favourable interactions between ligand molecules and the receptor molecule (Friesner et al., 2006). The XP (extra-precision) GlideScore algorithm was implemented to

achieve a much better prediction and ranking of compounds. Poses were generated as a zip file and are the combination of the position and orientation of a ligand relative to a receptor, alongside its conformation in flexible docking (https://drive.google.com/drive/folders/1q0j0zzz52mHphTI8L4OefLMWSCbUwQZC?usp=drive_link).

6.3 Results and Discussion

The docking results are sorted based on the best scoring pose for each compound (https://docs.google.com/spreadsheets/d/1xHrNoG6kJYowKf0PexvMuWEcq0gfr-Ny/edit?usp=drive_link&ouid=116684118916762575224&rtpof=true&sd=true).

The docking score represents the potential energy change that occurs when the protein and ligand interact (Li et al., 2019). A negative score indicates a strong binding, and a less negative or even positive score indicates a weak or non-existing binding (Narges et al., 2021). The higher the negative docking score values, the greater the ligand-protein interaction, i.e., it suggests that ligands with more negative scores rank higher. The summary result consists of the best docking scores of ligands for the 17 *S. aureus* biofilm-associated proteins. Table 2 revealed the protein-ligand interaction analysis, which is crucial for comprehending the mechanisms of the biological inhibition of biofilm formation in MDRSA. Also, they reveal a theoretical framework for the identification of novel antibiofilm agent and hastens the selection of drug hits and their subsequent developments. Figure 6.1 below represents the glide D-score distribution of the docked ligands with each of the 17 *S. aureus* biofilm-associated proteins that are important in the formation of biofilm in *S. aureus*. CNP0160461 and

CNP0037371 showed the strongest binding with all the docked biofilm-associated proteins.

Compound ID	DOCKING SCORES																
	2VR3	3ASW	3ATO	3AU0	4F1Z	4F2O	4F2A	4F27	5IQ6	7C7R	7VF0	7VFL	7VFM	7EC1	3GEU	7VFK	7VFN
CNP0023363	-7.404	-5.086	-6.177	-5.724	-5.944	-4.633	-5.548	-6.233	-5.713	-6.504	-4.744	-7.777	-5.578	-4.903	-4.533	-4.003	-4.717
CNP0037371	-8.967	-7.617	-7.667	-7.574	-8.424	-8.872	-8.427	-7.614	-7.337	-12.771	-10.77	-7.877	-7.67	-7.507	-7.583	-7.657	-7.893
CNP0081024	-3.343	-4.475	-6.216	-4.983	-4.671	-5.082	-5.246	-5.357	-5.676	-4.583	-4.832	-5.544	-6.132	-3.726	-4.411	-3.831	-4.044
CNP0082907	-4.381	-3.347	-4.627	-4.333	-4.93	-4.068	-4.148	-4.235	-3.775	-4.227	-3.734	-5.338	-3.614	-3.551	-3.306	-3.441	-3.236
CNP0085612	-7.062	-5.236	-5.051	-5.086	-5.418	-5.083	-5.404	-4.452	-5.193	-5.635	-4.882	-7.526	-3.963	-3.683	-4.113	-4.114	-4.489
CNP0126149	-6.18	-4.381	-5.464	-5.078	-4.881	-5.178	-4.798	-5.233	-4.728	-6.064	-5.685	-5.511	-5.67	-4.388	-4.298	-4.486	
CNP0160461	-12.356	-9.421	-8.954	-8.173	-9.173	-9.221	-8.967	-9.588	-10.11	-11.34	-7.542	-9.36	-7.196	-7.147	-6.337	-7.457	-9.753
CNP0161820	-4.926	-4.771	-5.08	-4.792	-4.425	-4.798	-5.435	-4.388	-5.267	-5.826	-5.721	-5.928	-5.908	-3.808	-4.473	-4.025	-3.935
CNP0189702	-8.715	-6.052	-7.507	-6.606	-6.093	-6.893	-5.95	-6.526	-5.477	-6.075	-6.648	-6.388	-5.263	-4.785	-4.881	-4.435	-5.668
CNP0221518	-4.471	-4.204	-5.063	-4.536	-4.25	-3.994	-4.842	-3.777	-4.68	-5.757	-3.595	-5.645	-3.965	-3.907	-3.743	-4.081	-5.101
CNP0259410	-4.567	-4.161	-4.281	-4.101	-4.871	-4.194	-5.164	-4.885	-4.797	-3.394	-3.61	-5.023	-3.766	-3.684	-3.834	-3.941	-3.967
CNP0282480	-8.764	-7.175	-6.462	-6.564	-7.227	-6.611	-8.076	-8.166	-6.567	-9.471	-6.771	-7.615	-6.218	-5.261	-5.805	-5.086	-7.658
CNP0326067	-5.186	-3.794	-4.684	-4.683	-3.711	-3.968	-4.634	-3.888	-3.836	-4.828	-3.841	-6.441	-2.743	-3.024	-3.231	-3.191	-3.521
CNP0357919	-8.126	-5.154	-6.464	-6.87	-5.328	-5.681	-6.314	-5.748	-5.073	-6.564	-5.602	-5.787	-6.801	-4.651	-5.044	-4.468	-5.143

Figure 6 Showing the glide docking scores of the *S. aureus* biofilm-associated protein with identified natural compounds from the consensus scoring of results from flexophore similarity chart and random forest predictive model. The coloured bars show how a score compares to others. Longer bars represent higher docking scores, shorter bars represent smaller docking scores, and missing values represent no existing docking interaction.

LIGAND	PROTEIN	INTERACTION(No of H-bond (residues))
CNP0037371	3GEU	5(TYR 96, SER 100, GLU 150)
CNP0037371	7C7R	8(LEU 720, TYR 526, GLY 380, PRO 379, ASP 378, GLU 539)
CNP0037371	7EC1	7(GLU 184, PRO 183, THR 203, ASN 7, VAL 226)
CNP0037371	7VFO	6(ASN 25, ASN 22, LYS 18, GLU 414, GLU 308, TYR 306)
CNP0037371	7VFK	8(ARG 150, GLU 161, LYS 174, GLU 202, VAL 185, PRO 183,TRP 41)
CNP0037371	7VFM	7(LEU 410, SER 409, TYR 393, SER 387, LEU 386, TYR 52)
CNP0160461	2VR3	8(LYS 389, THR 289, LYS 293, LEU 285, TYR 338, ASN 530)
CNP0160461	3ASW	7(LYS 333, LEU 301, ILE 300, ASN 268, ASP 330, GLY 269, ASP 272)
CNP0160461	3ATO	6(ASP 330, LYS 333, ILE 379, GLY 269)
CNP0160461	3AU0	7(ALA 448, THR 393, LYS 391, ARG 432, GLU 490)
CNP0160461	4F1Z	8(ASP 330, ASN 268, TYR 273, ASN 278, ILE 379)
CNP0160461	4F20	5(ALA 448, ILE 379, ARG 331)
CNP0160461	4F24	7(ASP 272, LYS 391, GLY 269, ASN 268, ILE 379, ASP 330)
CNP0160461	4F27	6(ASP 272, ARG 331, ASP 270, GLY 269, ILE 379)
CNP0160461	5JQ6	8(TYR 448, LEU 444, ALA 441, PRO 402, GLU 370, LYS 381)
CNP0160461	7VFL	10(GLH 406, LYS 334, HIE 246, SER 409, LEU410, ARG 329, LEU 13, LYS 18)
CNP0160461	7VFN	3(SER 394, ALA 396, GLN 466)

Table 2 Summary of interaction analysis for *S. aureus* biofilm-associated proteins and identified hits from consensus scoring

6.4 Conclusion

The docking study provided an opportunity to compare the top-ranked poses of the identified natural compounds based on consensus scoring. The ranking is a measure of the protein-ligand interactions of the natural compounds to identify potential antibiofilm compounds against multidrug-resistant *S. aureus*. The analysis of docking scores showed that the CNP0160461 and CNP0037371 have the strongest binding with identified *S. aureus* biofilm-associated proteins. Also, the result shows that the antibiofilm activity of the identified compounds against MDRSA is mediated by multiple targets i.e., the identified compounds are acting on multiple *S aureus* biofilm-associated proteins.

Chapter Seven

Conclusion, limitations, and recommendations

Multidrug-resistant *Staphylococcus aureus* poses threats to public health. Biofilm formation is the key virulence factor and a key survival strategy for *Staphylococcus aureus* and currently no approved drugs specifically targeting bacterial biofilms exist. Engineering next generation versions of current antibiotics results in substantially more failures than leads because it frequently involves screening large libraries, which are challenging to curate and fail to reflect the chemistry that is inherent to antibiotic molecules (Brown et al., 2014). Hence, this study was targeted at discovering natural compounds with antibiofilm activity against multidrug-resistant *S. aureus* using computational tools.

7.1 Summary of findings

The following is a succinct summary of the results for this study's objectives:

1. To collate and build a database of active and inactive antibiofilm compounds from bioassays in a literature search, collate the biofilm-associated proteins of S. aureus involved in cellular aggregation within the biofilm and retrieve compounds from natural compound databases.

Based on the search, a total of 323 active compounds and 105 inactive compounds were retrieved. A Biofilm structural database (BioSIM) was used to compile 39 *S. aureus* target proteins involved in biofilm formation. A query database of 411,180 natural compounds was collated from SANCDB and AfroDB.

2. To study the properties of compounds with reported antibiofilm properties e.g., compound flexophore and use them to query the natural compound database for active and inactive antibiofilm properties.

Flexophore similarity search approach detected compounds that could not be detected by simple ligand similarity search. From 411,180 natural compounds query dataset 43,596 compound pairs of compounds with flexophore similarity greater than 85%.

3. To build an antibiofilm predictive model for multidrug-resistant Staphylococcus aureus using a machine learning approach.

Machine learning approach utilised important molecular descriptors and features to build predictive models. The Random Forest and XGBOOST models were further used to predict the antibiofilm activity of compounds in the query databases. XGBOOST model result was more difficult to interpret, it predicted over 90 percent of the compounds in the query dataset as active with varying degrees of prediction confidence that is as low as 0.5. Random forest model result for activity prediction on the other hand was easy to interpret. A total of 30,097 compounds from 411,180 query dataset have Random Forest activity prediction confidence of greater than 0.85.

4. To perform consensus of potential antibiofilm compounds generated from ligand similarity searches and Machine Learning predictive model

Consensus scoring increases the likelihood of discovering the antibiofilm hit compounds by combining various scores to make up for flaws from individual

scoring functions. Consensus scoring identified 142 potential antibiofilm compounds.

5. To perform molecular docking studies on consensus scoring predictions of antibiofilm compounds with Staphylococcus aureus biofilm-associated proteins

Glide docking studies on the 142 selected compounds revealed possible mechanisms of action by analysing their interactions with *S. aureus* biofilm-associated proteins. CNP0160461 and CNP0037371 show the strongest binding among all the docked biofilm-associated proteins. This docking study shows that the antibiofilm activity of the identified compounds against MDRSA is by acting on multiple targets.

7.2 Limitations

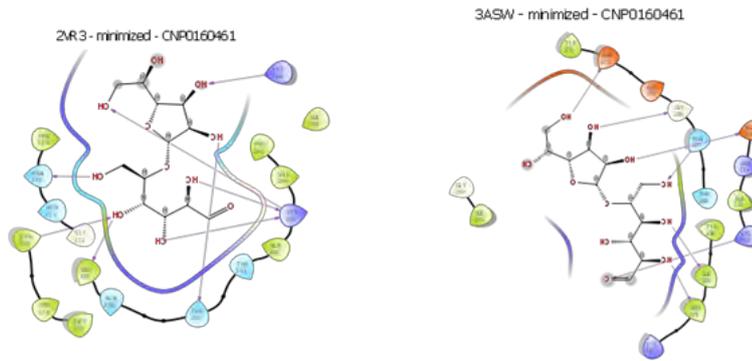
The results of ligand similarity searches were unable to unequivocally identify distinct differences in molecular structures of active and inactive compounds. Compounds that are structurally unrelated to known compounds with activity may nonetheless exhibit activity. In a recent study on a deep learning approach to finding new antibiotics, the model found antibacterial compounds that are structurally very different from known antibiotics (Stokes et al., 2020). Because algorithms can only learn from the data provided, the quality of the data used to train machine learning models has a significant impact on the accuracy of their predictions. The predictions made by machine learning models could be impacted if they learn unimportant features (noise) from the data provided. An interesting future work will be to discover how combining different classifiers in one model will result in prediction performance improvement.

7.3 Recommendations and Conclusion

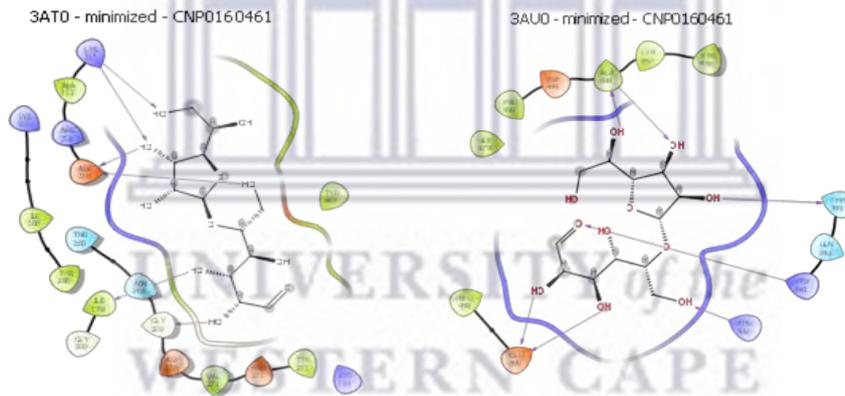
The compounds that were found to be active in this study have never been known to have antibiofilm activity against multidrug-resistant *S. aureus*. Hence, compounds with good binding affinity for biofilm-associated proteins of *Staphylococcus aureus* identified in this study could be researched further for the generation of potent antibiofilm agents for multidrug-resistant *Staphylococcus aureus*. Finally, to further support the *in-silico* predictions drawn from this study, extensive *in-vitro* and *in-vivo* studies are needed.



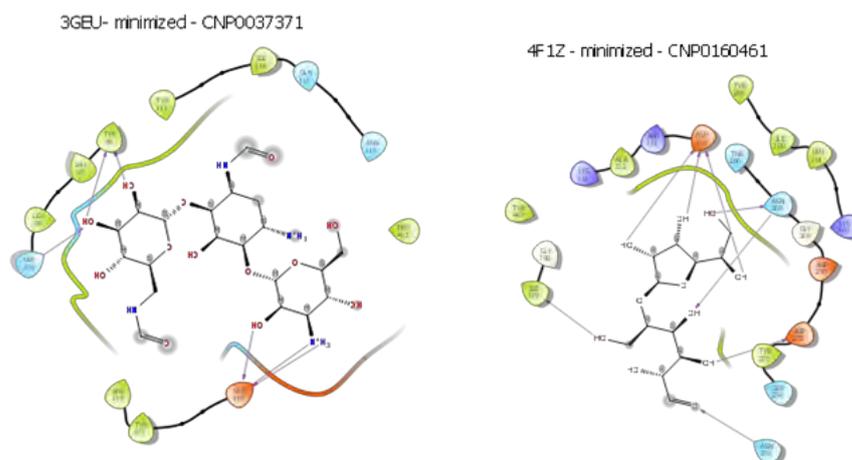
Appendices



Appendix (a) protein-ligand interaction for 2VR3 and 3ASW *Staphylococcus aureus* biofilm associated proteins (BaPs)

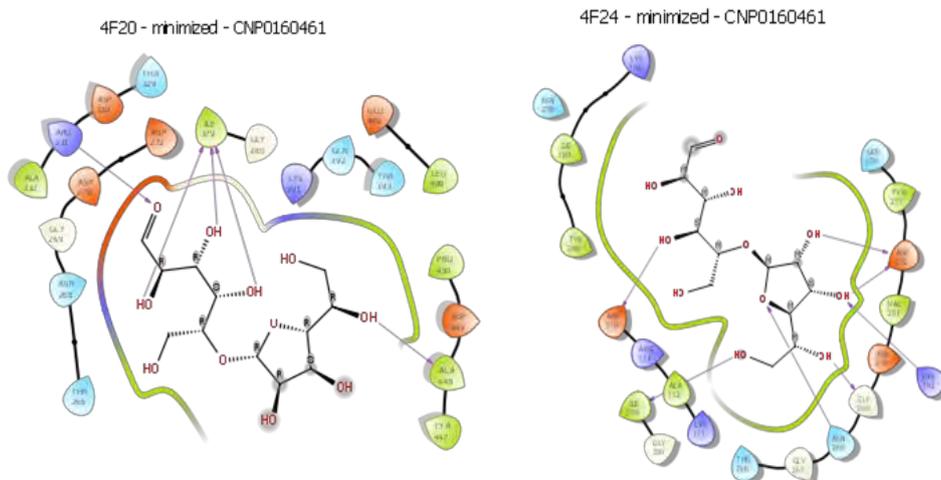


Appendix (b) protein-ligand interaction for 3AT0 and 3AU0 *Staphylococcus aureus* biofilm associated proteins (BAPs)

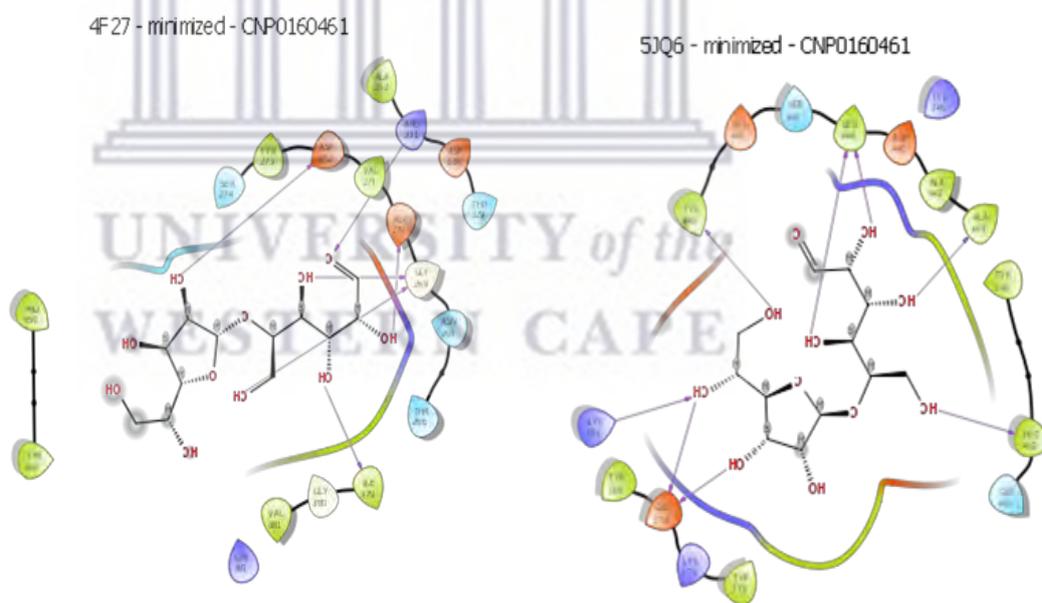


Appendix (c) protein-ligand interaction for 3GEU and 4F1Z *Staphylococcus aureus* biofilm associated proteins (BaPs)

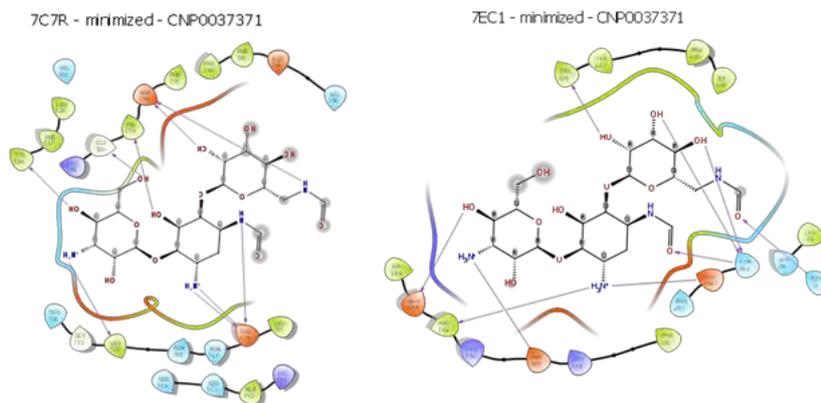




Appendix (d) protein-ligand interaction for 4F20 and 4F24 *Staphylococcus aureus* biofilm associated proteins (BaPs)

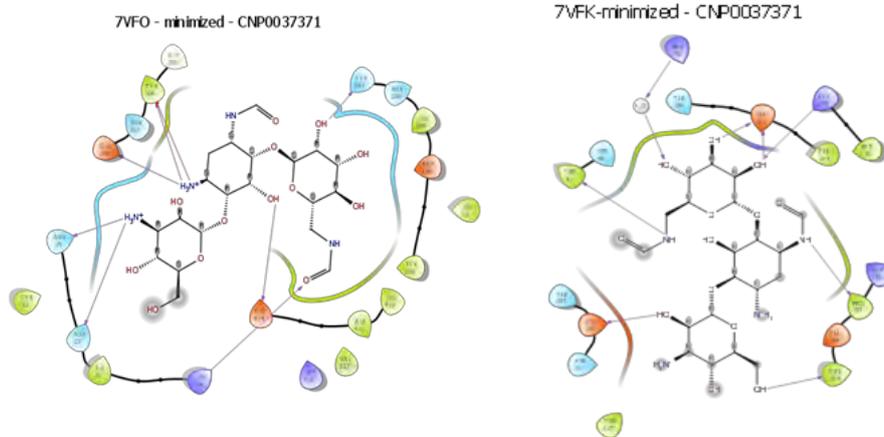


Appendix (e) protein-ligand interaction for 4F27 and 5JQ6 *Staphylococcus aureus* biofilm associated proteins (BaPs)

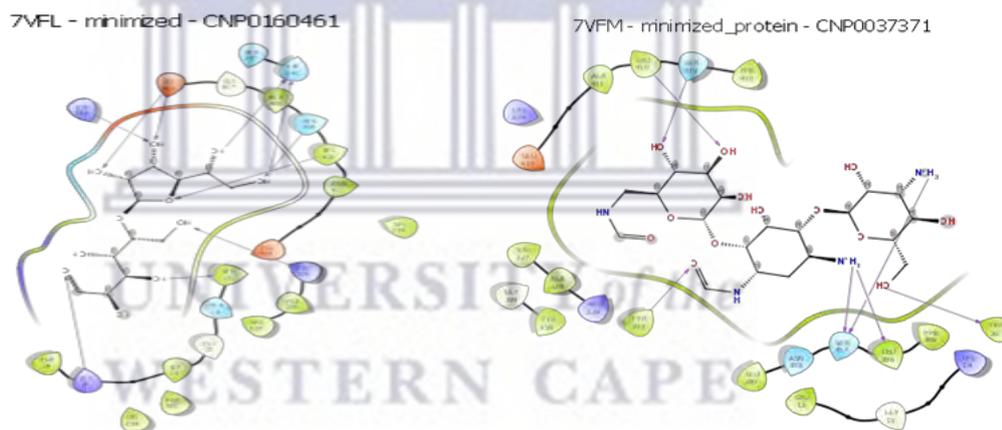


Appendix (f) protein-ligand interaction for 7C7R and 7EC1 *Staphylococcus aureus* biofilm associated proteins (BaPs)





Appendix (g) protein-ligand interaction for 7VFO and 7VFK *Staphylococcus aureus* biofilm associated proteins (BaPs)



Appendix (h) protein-ligand interaction for 7VFL and 7VFM *Staphylococcus aureus* biofilm associated proteins (BaPs)

References

- Abraham, N. M., & Jefferson, K. K. (2012). Staphylococcus aureus clumping factor B mediates biofilm formation in the absence of calcium. *Microbiology (United Kingdom)*, 158(6), 1504–1512. <https://doi.org/10.1099/mic.0.057018-0>
- Alves-Barroco, C., Roma-Rodrigues, C., Balasubramanian, N., Guimarães, M. A., Ferreira-Carvalho, B. T., Muthukumar, J., Nunes, D., Fortunato, E., Martins, R., Santos-Silva, T., Figueiredo, A. M. S., Fernandes, A. R., & Santos-Sanches, I. (2019). Biofilm development and computational screening for new putative inhibitors of a homolog of the regulatory protein BrpA in *Streptococcus dysgalactiae* subsp. *dysgalactiae*. *International Journal of Medical Microbiology*, 309(3-4), 169–181. <https://doi.org/10.1016/j.ijmm.2019.02.001>
- An, A. Y., Choi, K. Y. G., Baghela, A. S., & Hancock, R. E. W. (2021). An Overview of Biological and Computational Methods for Designing Mechanism-Informed Antibiofilm Agents. *Frontiers in Microbiology*, 12(April), 1–24. <https://doi.org/10.3389/fmicb.2021.640787>
- Archer, N. K., Mazaitis, M. J., William Costerton, J., Leid, J. G., Powers, M. E., & Shirtliff, M. E. (2011). Staphylococcus aureus biofilms: Properties, regulation and roles in human disease. *Virulence*, 2(5), 445–459. <https://doi.org/10.4161/viru.2.5.17724>
- Assis, L. M., Nedeljković, M., & Dessen, A. (2017). New strategies for targeting and treatment of multi-drug resistant Staphylococcus aureus. *Drug Resistance Updates*, 31, 1–14. <https://doi.org/10.1016/j.drug.2017.03.001>

Azeredo, J., & Sutherland, I. (2008). The Use of Phages for the Removal of Infectious Biofilms. *Current Pharmaceutical Biotechnology*, 9(4), 261–266.

<https://doi.org/10.2174/138920108785161604>

Bamford, N. C., Macphee, C. E., & Stanley-Wall, N. R. (2023). Microbial Primer: An introduction to biofilms – what they are, why they form and their impact on built and natural environments. *Microbiology (United Kingdom)*,

169(8), 1–6. <https://doi.org/10.1099/mic.0.001338>

Ban, T., Ohue, M., & Akiyama, Y. (2018). Multiple grid arrangement improves ligand docking with unknown binding sites: Application to the inverse docking problem. *Computational Biology and Chemistry*, 73, 139–146.

<https://doi.org/10.1016/j.compbiolchem.2018.02.008>

Bjarnsholt, T., Jensen, P. Ø., Rasmussen, T. B., Christophersen, L., Calum, H., Hentzer, M., Hougen, H. P., Rygaard, J., Moser, C., Eberl, L., Høiby, N., & Givskov, M. (2005). Garlic blocks quorum sensing and promotes rapid clearing of pulmonary *Pseudomonas aeruginosa* infections. *Microbiology*, 151(12), 3873–

3880. <https://doi.org/10.1099/mic.0.27955-0>

Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á., & Garcia-Fandino, R. (2023). The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. *Pharmaceuticals*,

16(6), 1–11. <https://doi.org/10.3390/ph16060891>

Carson, L., Gorman, S. P., & Gilmore, B. F. (2010). The use of lytic bacteriophages in the prevention and eradication of biofilms of *Proteus mirabilis*

and Escherichia coli. *FEMS Immunology and Medical Microbiology*, 59(3), 447–455. <https://doi.org/10.1111/j.1574-695X.2010.00696.x>

Cheung, G. Y. C., Bae, J. S., & Otto, M. (2021). Pathogenicity and virulence of *Staphylococcus aureus*. *Virulence*, 12(1), 547–569. <https://doi.org/10.1080/21505594.2021.1878688>

Chhabra, S., Kumar, S., & Parkesh, R. (2021). Chemical Space Exploration of DprE1 Inhibitors Using Chemoinformatics and Artificial Intelligence. *ACS Omega*, 6(22), 14430–14441. <https://doi.org/10.1021/acsomega.1c01314>

Chifiriuc, C., Grumezescu, V., Grumezescu, A. M., Saviuc, C., Lazăr, V., & Andronescu, E. (2012). Hybrid magnetite nanoparticles/ *Rosmarinus officinalis* essential oil nano-biosystem with antibiofilm activity. *Nanoscale Research Letters*, 7, 1–7. <https://doi.org/10.1186/1556-276X-7-209>

Coimbra, J. T. S., Feghali, R., Ribeiro, R. P., Ramos, M. J., & Fernandes, P. A. (2020). The importance of intramolecular hydrogen bonds on the translocation of the small drug piracetam through a lipid bilayer. *RSC Advances*, 11(2), 899–908. <https://doi.org/10.1039/d0ra09995c>

Cortés-Ciriano, I., Škuta, C., Bender, A., & Svozil, D. (2020). QSAR-derived affinity fingerprints (part 2): Modelling performance for potency prediction. *Journal of Cheminformatics*, 12(1), 1–17. <https://doi.org/10.1186/s13321-020-00444-5>

Dalman, M., Bhatta, S., Nagajothi, N., Thapaliya, D., Olson, H., Naimi, H. M., & Smith, T. C. (2019). Characterizing the molecular epidemiology of

Staphylococcus aureus across and within fitness facility types. *BMC Infectious Diseases*, 19(1), 1–10. <https://doi.org/10.1186/s12879-019-3699-7>

Danishuddin, & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: *paradigm for drug design*. *Drug Discovery Today*, 21(8), 1291–1302. <https://doi.org/10.1016/j.drudis.2016.06.013>

Dasilva, L., Finer, Y., Friedman, S., Basrani, B., & Kishen, A. (2013). Biofilm formation within the interface of bovine root dentin treated with conjugated chitosan and sealer containing chitosan nanoparticles. *Journal of Endodontics*, 39(2), 249–253. <https://doi.org/10.1016/j.joen.2012.11.008>

Defoirdt, T., Brackman, G., & Coenye, T. (2013). Quorum sensing inhibitors: How strong is the evidence? *Trends in Microbiology*, 21(12), 619–624. <https://doi.org/10.1016/j.tim.2013.09.006>

Divakar, S., Lama, M., & Asad U., K. (2019). Antibiotics versus biofilm: an emerging battleground in microbial communities | *Enhanced Reader*. *Antimicrobial Resistance and Infection Control*, 3, 1–10. <https://doi.org/10.1186/s13756-019-0533-3>

Doman, T. N., McGovern, S. L., Witherbee, B. J., Kasten, T. P., Kurumbail, R., Stallings, W. C., Connolly, D. T., & Shoichet, B. K. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of Medicinal Chemistry*, 45(11), 2213–2221. <https://doi.org/10.1021/jm010548w>

- Drie, J. H. (2007). Computer-aided drug design: The next 20 years. *Journal of Computer-Aided Molecular Design*, 21(10–11), 591–601. <https://doi.org/10.1007/s10822-007-9142-y>
- Duarte, S., Gregoire, S., Singh, A. P., Vorsa, N., Schaich, K., Bowen, W. H., & Koo, H. (2006). Inhibitory effects of cranberry polyphenols on formation and acidogenicity of *Streptococcus mutans* biofilms. *FEMS Microbiology Letters*, 257(1), 50–56. <https://doi.org/10.1111/j.1574-6968.2006.00147.x>
- Egıeyeh S, Syce J, Malan S.F, Christoffels A. (2018) Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. *PLoS ONE* 13(9): e0204644. <https://doi.org/10.1371/journal>.
- Elaziz, M. A., Moemen, Y. S., Hassanien, A. E., & Xiong, S. (2018). Quantitative Structure-Activity Relationship Model for HCVNS5B inhibitors based on an Antlion Optimizer-Adaptive Neuro-Fuzzy Inference System. *Scientific Reports*, 8(1), 1–17. <https://doi.org/10.1038/s41598-017-19122-y>
- Ezeh, C. K., Eze, C. N., Dibua, M. E. U., & Emencheta, S. C. (2023). A meta-analysis on the prevalence of resistance of *Staphylococcus aureus* to different antibiotics in Nigeria. *Antimicrobial Resistance and Infection Control*, 12(1), 1–22. <https://doi.org/10.1186/s13756-023-01243-x>
- Foster, T. J. (2016). The remarkably multifunctional fibronectin-binding proteins of *Staphylococcus aureus*. *European Journal of Clinical Microbiology and Infectious Diseases*, 35(12), 1923–1931. <https://doi.org/10.1007/s10096-016-2763-0>

- Foster, T. J., Geoghegan, J. A., Ganesh, V. K., & Höök, M. (2014). Adhesion, invasion, and evasion: The many functions of the surface proteins of *Staphylococcus aureus*. *Nature Reviews Microbiology*, 12(1), 49–62. <https://doi.org/10.1038/nrmicro3161>
- Friesner, R.A., Murphy, R.B., Repasky M.P., Frye, L., Greenwood J.R., Halgren T.A., Sanschagrin P.C. & Mainz D.T. (2006) Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21), 6177-6196. <https://doi.org/10.1021/jm051256o>
- Gillet, V. J. (2019). Applications of Chemoinformatic in Drug Discovery. *Biomolecular and Bioanalytical Techniques*, 17–36. <https://doi.org/10.1002/9781119483977.ch2>
- Groome, M. J., Albrich, W. C., Wadula, J., Khoosal, M., & Madhi, S. A. (2012). Community-onset *Staphylococcus aureus* bacteraemia in hospitalised African children: High incidence in HIV-infected children and high prevalence of multidrug resistance. *Paediatrics and International Child Health*, 32(3), 140–146. <https://doi.org/10.1179/1465328111Y.0000000044>
- Guo, Y., Song, G., Sun, M., Wang, J., & Wang, Y. (2020). Prevalence and Therapies of Antibiotic-Resistance in *Staphylococcus aureus*. *Frontiers in Cellular and Infection Microbiology*, 10(March), 1–11. <https://doi.org/10.3389/fcimb.2020.00107>

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modelling*, 49(2), 377–389.

<https://doi.org/10.1021/ci800324m>

Hawas, S., Verderosa, A. D., & Totsika, M. (2022). Combination Therapies for Biofilm Inhibition and Eradication: A Comparative Review of Laboratory and Preclinical Studies. *Frontiers in Cellular and Infection Microbiology*, 12(February), 1–19. <https://doi.org/10.3389/fcimb.2022.850030>

Helguera, A., Combes, R., Gonzalez, M., & Cordeiro, M. N. (2008). Applications of 2D Descriptors in Drug Design: A DRAGON Tale. *Current Topics in Medicinal Chemistry*, 8(18), 1628–1655.

<https://doi.org/10.2174/156802608786786598>

Hewagama, S., Spelman, T., & Einsiedel, L. J. (2012). Staphylococcus aureus bacteraemia at Alice Springs Hospital, Central Australia, 2003-2006. *Internal Medicine Journal*, 42(5), 505–512. <https://doi.org/10.1111/j.1445-5994.2011.02449.x>

Huh, A. J., & Kwon, Y. J. (2011). “Nano-antibiotics”: A new paradigm for treating infectious diseases using nanomaterials in the antibiotics resistant era. *Journal of Controlled Release*, 156(2), 128–145.

<https://doi.org/10.1016/j.jconrel.2011.07.002>

Id, A. M., Hulst, J. M., Boon, M., Witters, P., Fernandez-Illatas, C., Asseiceira, I., Calvo-lerma, J., Basagoiti, I., Traver, V., Boeck, K. De, & Ribes-koninckx, C.

(2018). *Optimisation of children's z-score calculation based on new statistical techniques*. 1–13. <https://doi.org/10.1371/journal.pone.0208362>

Jaśkiewicz, M., Janczura, A., Nowicka, J., & Kamysz, W. (2019). Methods used for the eradication of staphylococcal biofilms. *Antibiotics*, 8(4). <https://doi.org/10.3390/antibiotics8040174>

Jernigan, J. A., Hatfield, K. M., Wolford, H., Nelson, R. E., Olubajo, B., Reddy, S. C., McCarthy, N., Paul, P., McDonald, L. C., Kallen, A., Fiore, A., Craig, M., & Baggs, J. (2020). Multidrug-Resistant Bacterial Infections in U.S. Hospitalized Patients, 2012–2017. *New England Journal of Medicine*, 382(14), 1309–1319. <https://doi.org/10.1056/nejmoa1914433>

Joris, F., Manshian, B. B., Peynshaert, K., De Smedt, S. C., Braeckmans, K., & Soenen, S. J. (2013). Assessing nanoparticle toxicity in cell-based assays: Influence of cell culture parameters and optimized models for bridging the in vitro-in vivo gap. *Chemical Society Reviews*, 42(21), 8339–8359. <https://doi.org/10.1039/c3cs60145e>

Kakoullis, L., Papachristodoulou, E., Chra, P., & Panos, G. (2021). Mechanisms of antibiotic resistance in important gram-positive and gram-negative pathogens and novel antibiotic solutions. *Antibiotics*, 10(4). <https://doi.org/10.3390/antibiotics10040415>

Kalyaanamoorthy, S., & Chen, Y. P. P. (2011). Structure-based drug design to augment hit discovery. *Drug Discovery Today*, 16(17–18), 831–839. <https://doi.org/10.1016/j.drudis.2011.07.006>

- Khedkar, S., Malde, A., Coutinho, E., & Srivastava, S. (2007). Pharmacophore Modeling in Drug Discovery and Development: An Overview. *Medicinal Chemistry*, 3(2), 187–197. <https://doi.org/10.2174/157340607780059521>
- Kim, S. W., Chang, I. M., & Oh, K. B. (2002). Inhibition of the Bacterial Surface Protein Anchoring Transpeptidase Sortase by Medicinal Plants. *Bioscience, Biotechnology and Biochemistry*, 66(12), 2751–2754. <https://doi.org/10.1271/bbb.66.2751>
- Klevens, R. M., Morrison, M. A., Nadle, J., Petit, S., Gershman, K., Ray, S., Harrison, L. H., Lynfield, R., Dumyati, G., Townes, J. M., Craig, A. S., Zell, E. R., Fosheim, G. E., McDougal, L. K., Carey, R. B., & Fridkin, S. K. (2007). Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. *Journal of the American Medical Association*, 298(15), 1763–1771. <https://doi.org/10.1001/jama.298.15.1763>
- Koo, H., Allan, R. N., Howlin, R. P., Hall-Stoodley, L., & Stoodley, P. (2018). Targeting microbial biofilms: current and prospective therapeutic strategies. *Physiology & Behavior*, 176(1), 139–148. <https://doi.org/10.1038/nrmicro.2017.99>
- Kranjec, C., Angeles, D. M., Mårli, M. T., Fernández, L., García, P., Kjos, M., & Diep, D. B. (2021). Staphylococcal biofilms: Challenges and novel therapeutic perspectives. *Antibiotics*, 10(2), 1–30. <https://doi.org/10.3390/antibiotics10020131>

- Lazar, V., Holban, A. M., Curutiu, C., & Chifiriuc, M. C. (2021). Modulation of Quorum Sensing and Biofilms in Less Investigated Gram-Negative ESKAPE Pathogens. *Frontiers in Microbiology*, 12(July), 1–18. <https://doi.org/10.3389/fmicb.2021.676510>
- Li, X. H., & Lee, J. H. (2017). Antibiofilm agents: A new perspective for antimicrobial strategy. *Journal of Microbiology*, 55(10), 753–766. <https://doi.org/10.1007/s12275-017-7274-x>
- Lister, J. L., & Horswill, A. R. (2014). Staphylococcus aureus biofilms: Recent developments in biofilm dispersal. *Frontiers in Cellular and Infection Microbiology*, 4(DEC), 1–9. <https://doi.org/10.3389/fcimb.2014.00178>
- Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8), 1538–1546. <https://doi.org/10.1016/j.drudis.2018.05.010>
- Lu, L., Hu, W., Tian, Z., Yuan, D., Yi, G., Zhou, Y., Cheng, Q., Zhu, J., & Li, M. (2019). Developing natural products as potential anti-biofilm agents. *Chinese Medicine (United Kingdom)*, 14(1), 1–17. <https://doi.org/10.1186/s13020-019-0232-2>
- Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3), 221–234. <https://doi.org/10.1007/s10822-013-9644-8>

Maggiora, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), 3186–3204. <https://doi.org/10.1021/jm401411z>

Marvasi, M., Chen, C., Carrazana, M., Durie, I. A., & Teplitski, M. (2014). Systematic analysis of the ability of Nitric Oxide donors to dislodge biofilms formed by *Salmonella enterica* and *Escherichia coli* O157:H7. *AMB Express*, 4(1), 1–11. <https://doi.org/10.1186/s13568-014-0042-y>

Masák, J., Čejková, A., Schreiberová, O., & Řezanka, T. (2014). *Pseudomonas* biofilms: Possibilities of their control. *FEMS Microbiology Ecology*, 89(1), 1–14. <https://doi.org/10.1111/1574-6941.12344>

Mathur, S., & Hoskins, C. (2017). Drug development: Lessons from nature (Review). *Biomedical Reports*, 6, 612–614. <https://doi.org/10.3892/br.2017.909>

Michel, M., Visnes, T., Homan, E. J., Seashore-Ludlow, B., Hedenström, M., Wiita, E., Vallin, K., Paulin, C. B. J., Zhang, J., Wallner, O., Scobie, M., Schmidt, A., Jenmalm-Jensen, A., Warpman Berglund, U., & Helleday, T. (2019). Computational and Experimental Druggability Assessment of Human DNA Glycosylases. *ACS Omega*, 4(7), 11642–11656. <https://doi.org/10.1021/acsomega.9b00162>

Montanaro, L., Poggi, A., Visai, L., Ravaioli, S., Campoccia, D., Speziale, P., & Arciola, C. R. (2011). Extracellular DNA in biofilms. *International Journal of Artificial Organs*, 34(9), 824–831. <https://doi.org/10.5301/ijao.5000051>

Moormeier, D. E., & Bayles, K. W. (2017). *Staphylococcus aureus* biofilm: a complex developmental organism. *Molecular Microbiology*, 104(3), 365–376. <https://doi.org/10.1111/mmi.13634>

Nantasenamat, C., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2010). Advances in computational methods to predict the biological activity of compounds. *Expert Opinion on Drug Discovery*, 5(7), 633–654. <https://doi.org/10.1517/17460441.2010.492827>

Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., & Andrade, C. H. (2018). QSAR-based virtual screening: Advances and applications in drug discovery. *Frontiers in Pharmacology*, 9(NOV), 1–7. <https://doi.org/10.3389/fphar.2018.01275>

Nhat, D., Darren, P., Subhagata, R. F., & Amit, C. (2023). Towards Effective Consensus Scoring in Structure Based Virtual Screening. *Interdisciplinary Sciences: Computational Life Sciences*, 15(1), 131–145. <https://doi.org/10.1007/s12539-022-00546-8>

O’Loughlin, C. T., Miller, L. C., Siryaporn, A., Drescher, K., Semmelhack, M. F., & Bassler, B. L. (2013). A quorum-sensing inhibitor blocks *Pseudomonas aeruginosa* virulence and biofilm formation. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44), 17981–17986. <https://doi.org/10.1073/pnas.1316981110>

Opdensteinen, P., Dietz, S. J., Gengenbach, B. B., & Buyel, J. F. (2021). Expression of Biofilm-Degrading Enzymes in Plants and Automated High-

- Throughput Activity Screening Using Experimental *Bacillus subtilis* Biofilms. *Frontiers in Bioengineering and Biotechnology*, 9(September), 1–12. <https://doi.org/10.3389/fbioe.2021.708150>
- Otto, M. (2019). Staphylococcal biofilms. *Gram-Positive Pathogens*, 699–711. <https://doi.org/10.1128/9781683670131.ch43>
- Paraszkiewicz, K., Moryl, M., Płaza, G., Bhagat, D., K. Satpute, S., & Bernat, P. (2021). Surfactants of microbial origin as antibiofilm agents. *International Journal of Environmental Health Research*, 31(4), 401–420. <https://doi.org/10.1080/09603123.2019.1664729>
- Paraszkiewicz, K., Moryl, M., Płaza, G., Bhagat, D., K. Satpute, S., & Bernat, P. (2021). Surfactants of microbial origin as antibiofilm agents. *International Journal of Environmental Health Research*, 31(4), 401–420. <https://doi.org/10.1080/09603123.2019.1664729>
- Penesyanyan, A., Nagy, S. S., Kjelleberg, S., Gillings, M. R., & Paulsen, I. T. (2019). Rapid microevolution of biofilm cells in response to antibiotics. *Npj Biofilms and Microbiomes*, 5(1). <https://doi.org/10.1038/s41522-019-0108-3>
- Peter, S. C., Dhanjal, J. K., Malik, V., Radhakrishnan, N., Jayakanthan, M., Sundar, D., & Sundar, D. (2018). Quantitative structure-activity relationship (QSAR): Modelling approaches to biological applications. In *Encyclopaedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3, (January). <https://doi.org/10.1016/B978-0-12-809633-8.20197-0>

Rabin, N., Zheng, Y., Opoku-Temeng, C., Du, Y., Bonsu, E., & Sintim, H. O. (2015). Agents that inhibit bacterial biofilm formation. *Future Medicinal Chemistry*, 7(5), 647–671. <https://doi.org/10.4155/fmc.15.7>

Rajput, A., Thakur, A., Sharma, S., & Kumar, M. (2018). ABiofilm: A resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Research*, 46(D1), D894–D900. <https://doi.org/10.1093/nar/gkx1157>

Sahoo, A., Swain, S. S., Behera, A., Sahoo, G., Mahapatra, P. K., & Panda, S. K. (2021). Antimicrobial peptides derived from insects offer a novel therapeutic option to combat biofilm: A Review. *Frontiers in Microbiology*, 12(June). <https://doi.org/10.3389/fmicb.2021.661195>

Saising, J., Ongsakul, M., & Voravuthikunchai, S. P. (2011). *Rhodomyrtus tomentosa* (Aiton) Hassk. ethanol extract and rhodomyrtone: A potential strategy for the treatment of biofilm-forming staphylococci. *Journal of Medical Microbiology*, 60(12), 1793–1800. <https://doi.org/10.1099/jmm.0.033092-0>

Sanchez, G. (2013) Protein-Ligand docking: Current and future challenges. *PROTEINS: Structure, Function, and Bioinformatics* 65, 15–26. <https://doi.org/10.1002/prot>

Sander, T., Freyss, J., Von Korff, M., & Rufener, C. (2015). DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modelling*, 55(2), 460–473. <https://doi.org/10.1021/ci500588j>

Sathyannarayanan, M. B., Balachandranath, R., Genji Srinivasulu, Y., Kannaiyan, S. K., & Subbiahdoss, G. (2013). The Effect of Gold and Iron-Oxide Nanoparticles on Biofilm-Forming Pathogens. *Microbiology*, 2013, 1–5. <https://doi.org/10.1155/2013/272086>

Sato, A., Yamaguchi, T., Hamada, M., Ono, D., Sonoda, S., Oshiro, T., et al. (2019). Morphological and biological characteristics of *Staphylococcus aureus* biofilm formed in the presence of plasma. *Microb. Drug. Resist.* 25, 668–676. <http://doi:10.1089/mdr.2019.0068>

Satpute, S. K., Kulkarni, G. R., Banpurkar, A. G., Banat, I. M., Mone, N. S., Patil, R. H., & Cameotra, S. S. (2016). Biosurfactant/s from *Lactobacilli* species: Properties, challenges and potential biomedical applications. *Journal of Basic Microbiology*, 56(11), 1140–1158. <https://doi.org/10.1002/jobm.201600143>

Saurabh Pal, R. A. (2021). Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(6), 2650–2665. <https://doi.org/10.17762/turcomat.v12i6.5765>

Schaumburg, F., Alabi, A. S., Peters, G., & Becker, K. (2014). New epidemiology of *Staphylococcus aureus* infection in Africa. In *Clinical Microbiology and Infection* 20(7), 589–596. <https://doi.org/10.1111/1469-0691.12690>

Schuffenhauer, A. (2012). Computational methods for scaffold hopping. Wiley Interdisciplinary Reviews: *Computational Molecular Science*, 2(6), 842–867. <https://doi.org/10.1002/wcms.1106>

Sigaúque, B., Roca, A., Mandomando, I., Morais, L., Quintó, L., Sacarlal, J., MacEte, E., Nhamposa, T., MacHevo, S., Aide, P., Bassat, Q., Bardaji, A., Nhalungo, D., Soriano-Gabarró, M., Flannery, B., Menendez, C., Levine, M. M., & Alonso, P. L. (2009). Community-acquired bacteremia among children admitted to a rural hospital in Mozambique. *Paediatric Infectious Disease Journal*, 28(2), 108–113. <https://doi.org/10.1097/INF.0b013e318187a87d>

Singh, B. N., Prateeksha, Upreti, D. K., Singh, B. R., Defoirdt, T., Gupta, V. K., De Souza, A. O., Singh, H. B., Barreira, J. C. M., Ferreira, I. C. F. R., & Vahabi, K. (2017). Bactericidal, quorum quenching and antibiofilm nano-factories: a new niche for nanotechnologists. *Critical Reviews in Biotechnology*, 37(4), 525–540. <https://doi.org/10.1080/07388551.2016.1199010>

Sinsomboonthong, S. (2022). Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification. *International Journal of Mathematics and Mathematical Sciences*, 2022. <https://doi.org/10.1155/2022/3584406>

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological Reviews*, 66(1), 334–395. <https://doi.org/10.1124/pr.112.007336>

Steinig, E. J., Duchene, S., Robinson, D. A., Monecke, S., Yokoyama, M., Laabei, M., Slickers, P., Andersson, P., Williamson, D., Kearns, A., Goering, R. V, Dickson, E., Shore, A. C., Coleman, D. C., Pantosti, A., Lencastre, H. De, Westh, H., Kobayashi, N., Heffernan, H., & Tong, Y. C. (2019). *Staphylococcus aureus* lineage from the Indian Subcontinent. *Clinical science and epidemiology* 10(6), 1–20. <https://doi.org/10.1128/mBio.01105-19>.

Stumpfe, D., Hu, Y., Dimova, D., & Bajorath, J. (2014). Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(1), 18–28. <https://doi.org/10.1021/jm401120g>

Tamma, P. D., Cosgrove, S. E., & Maragakis, L. L. (2012). Combination therapy for treatment of infections with gram-negative bacteria. *Clinical Microbiology Reviews*, 25(3), 450–470. <https://doi.org/10.1128/CMR.05041-11>

Tayeb-fligelman, E., Tabachnikov, O., Moshe, A., Goldshmidt-tran, O., Sawaya, M. R., Coquelle, N., Colletier, J., & Landau, M. (2017). The cytotoxic *Staphylococcus aureus*. *Science (New York, N.Y.)*, 355(6327), 21–24. <http://www.ncbi.nlm.nih.gov/pubmed/28232575>

Thukkaram, M., Sitaram, S., Kannaiyan, S. K., & Subbiahdoss, G. (2014). Antibacterial efficacy of iron-oxide nanoparticles against biofilms on different biomaterial surfaces. *International Journal of Biomaterials*, 2014. 1-6. <https://doi.org/10.1155/2014/716080>

Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L., & Fowler, V. G. (2015). *Staphylococcus aureus* infections: Epidemiology, pathophysiology,

clinical manifestations, and management. *Clinical Microbiology Reviews*, 28(3), 603–661. <https://doi.org/10.1128/CMR.00134-14>

Tu, Y. (2019). In EEG Signal Processing and Feature Extraction. *Machine learning*. 301-319. https://doi.org/10.1007/978-981-13-9113-2_15

Verderosa, A. D., Totsika, M., & Fairfull-Smith, K. E. (2019). Bacterial Biofilm Eradication Agents: A Current Review. *Frontiers in Chemistry*, 7(November), 1–17. <https://doi.org/10.3389/fchem.2019.00824>

Vestby, L. K., Grønseth, T., Simm, R., & Nesse, L. L. (2020). Bacterial biofilm and its role in the pathogenesis of disease. *Antibiotics*, 9(2). 1-29. <https://doi.org/10.3390/antibiotics9020059>

Von Korff, M., Freyss, J., & Sander, T. (2008). Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *Journal of Chemical Information and Modeling*, 48(4), 797–810. <https://doi.org/10.1021/ci700359j>

Zhang, R., Li, X., Zhang, X., Qin, H., & Xiao, W. (2021). Machine learning approaches for elucidating the biological effects of natural products. *Natural Product Reports*, 38(2), 346–361. <https://doi.org/10.1039/d0np00043d>



UNIVERSITY *of the*
WESTERN CAPE